

From the Department of Medical Biochemistry and Biophysics  
Karolinska Institutet, Stockholm, Sweden

# **COMPUTATIONAL EXPLORATION OF THE MEDIUM-CHAIN DEHYDROGENASE / REDUCTASE SUPERFAMILY**

Linus J. Östberg



**Karolinska  
Institutet**

Stockholm 2017

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet.

Printed by E-Print AB 2017

©Linus J. Östberg, 2017

ISBN 978-91-7676-650-7

# COMPUTATIONAL EXPLORATION OF THE MEDIUM-CHAIN DEHYDROGENASE / REDUCTASE SUPERFAMILY

## THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

**Linus J. Östberg**

*Principal Supervisor:*

Prof. Jan-Olov Höög  
Karolinska Institutet  
Department of Medical Biochemistry and  
Biophysics

*Co-supervisor:*

Prof. Bengt Persson  
Uppsala University  
Department of Cell and Molecular Biology

*Opponent:*

Prof. Jaume Farrés  
Autonomous University of Barcelona  
Department of Biochemistry and  
Molecular Biology

*Examination Board:*

Prof. Erik Lindahl  
Stockholm University  
Department of Biochemistry and Biophysics

Assoc. Prof. Tanja Slotte  
Stockholm University  
Department of Ecology, Environment  
and Plant Sciences

Prof. Elias Arnér  
Karolinska Institutet  
Department of Medical Biochemistry and  
Biophysics





## ABSTRACT

The medium-chain dehydrogenase/reductase (MDR) superfamily is a protein family with more than 170,000 members across all phylogenetic branches. In humans there are 18 representatives. The entire MDR superfamily contains many protein families such as alcohol dehydrogenase, which in mammals is in turn divided into six classes, class I–VI (ADH1–6). Most MDRs have enzymatical functions, catalysing the conversion of alcohols to aldehyde/ketones and vice versa, but the function of some members is still unknown.

In the first project, a methodology for identifying and automating the classification of mammalian ADHs was developed using BLAST for identification and class-specific hidden Markov models were generated for identification. By using the developed methodology, multiple new mammalian ADH members were identified. Finally, the generation of a phylogenetic tree of the sequences showed the existence of a sixth class, ADH6, in most non-primate mammals, though the sequences are commonly misclassified as ADH5 or ADH1-like in the sequence databases.

The second project focused on the study of mammalian ADH5, which has never been isolated as a native protein, and whose function is unknown. The first part of the project was the expression of ADH5 fusion proteins in *E. coli* and COS cells (human ADH5 with glutathione-S-transferase in *E. coli* and rat ADH5-green fluorescent protein in COS cells). The proteins were expressed, but had no activity with any traditional ADH substrates. The results also indicated potential problems with the stability of the protein.

The continuation of the project was the analysis of the structure using computational methods. A structural model of ADH5 was generated using the homolog ADH1C as template. Molecular dynamics was subsequently used to study the properties of the model. Along with the structural analysis, extensive sequence analysis was also performed, identifying multiple positions that were unique for ADH5, e.g. Lys51 at the active site and Gly305 in the dimer-interacting region, which replace a highly conserved Pro found in ADH1–4. The combination of the structural simulations and the sequence analysis led to the conclusion that the lack of success in the isolation of ADH5 could potentially be explained by instabilities in the region involved in dimer formation, preventing the formation of the active dimers found in other ADHs. The function of ADH5 is therefore concluded to be different than that of other ADHs, but is as of yet still unknown.

The final project focused on the study of a set of human MDRs, using a combination of analyses of the structures and sequences, leading to the development of theoretical models of the binding pockets in each of the proteins, pinpointing the important residues. The positions identified to be involved in the binding of the coenzyme NADP(H) were similar between the proteins and matched currently available information in the databases, as well as further residues. The residues involved in the binding of substrates varied between the proteins, and the analysis led to the identification of three different types of substrate binding.



# POPULAR SCIENCE SUMMARY

Back in 2003, the human genome was finally sequenced. With all this information in hand, we know the basic code for how humans work. Now we need to understand it.

The genome (the DNA) encodes proteins, which in turn build up the major part of our bodies, from the smallest strand of hair to our brains. We know how many of these proteins work, but at the same time there are many others for which we do not know the function.

The work presented in this thesis focuses on the medium-chain dehydrogenase/reductase (MDR) proteins, which form a large family of enzymes that is represented by 18 different members in humans. The function of many of these proteins are known, but the function of some of them still needs to be investigated.

Two of the projects presented here focus on the functions of the alcohol dehydrogenase (ADH) protein family, which is part of the MDR family and correspond to seven of the 18 human MDR members. As the name implies, the ADHs functions on alcohols, changing them to aldehydes, and the family also includes the enzyme responsible for the breakdown of ethanol to acetaldehyde.

The first project is focused on the classification (naming) of the mammalian ADH members (paper II). In humans, only five of the six ADH classes are present, and they function on different alcohols. In the databases containing protein sequences, the classification of them is not always matching reality, and the project developed a methodology to identify ADHs in the databases and confirm (or fix) their name to match the correct class. As the methodology that was developed also led to the gathering of a high number of mammalian ADHs, it was also possible to do some comparisons between the different species, e.g. which species had which classes of ADH.

The second project focused on the analysis of the class V ADH, ADH5 (paper I and III). The function of ADH5 is, as opposed to the other ADHs found in humans, unknown, and nobody has ever been able to study the protein in a lab. In the first part, an attempt was made to isolate and study ADH5 by merging ADH5 with other proteins, first glutathione-S-transferase and then green fluorescent protein. The merged proteins could be isolated, but they did not act on any of the expected molecules (that are known to work with the other ADHs). It was also impossible to isolate the proteins without the extra part that had been added.

In the second part, the focus was on using bioinformatics (computational methods) to simulate and analyse the properties of ADH5. A model of the structure was made using the other human ADHs as templates, and then the model was analysed using molecular dynamics (a method that simulates the movements of the atoms over time). Along with this, the method from the first project was used to identify the sequence of the ADH5 protein in as many mammals as possible. By studying the sequences and the simulations, it was seen that ADH5 without a doubt is related to the other ADHs, but is a bit of an odd sibling. At many positions where the other ADHs had the same amino acid, ADH5 sometimes had something else. It was also seen that the structure of ADH5 was not behaving in the same way as the other ADHs. As such, it probably has a function which is different (but potentially related) to that of the other ADHs.

The final project (paper IV) focused on nine other human MDRs that were not part of the ADH protein family. The aim was to analyse the proteins, with focus on the area where they bind other molecules in order to modify them. The nine proteins were studied by identifying sequences in a similar way to what was used to find ADHs in the first project, followed by a sequence analysis similar to what was used to study ADH5, and finally some new ways of analysis as well. From these analyses, it could be seen that the non-ADH MDRs have at least three different types of substrate binding (where the modification of other molecules occur), using different amino acids for their effects. Further, it could be seen that NADP(H), which is used as a coenzyme (helper molecule for the enzyme) to do the modifications to other molecules) binds in a very similar way in all nine MDRs.

## LIST OF PUBLICATIONS

- I **Östberg LJ**, Strömberg P, Hedberg JJ, Persson B, Höög J-O. Analysis of mammalian alcohol dehydrogenase 5 (ADH5): Characterisation of rat ADH5 with comparisons to the corresponding human variant. *Chem Biol Interact.* 2013;202(1–3):97–103.
- II **Östberg LJ**, Persson B, Höög J-O. The mammalian alcohol dehydrogenase genome shows several gene duplications and gene losses resulting in a large set of different enzymes including pseudoenzymes. *Chem Biol Interact.* 2015;234:80–84.
- III **Östberg LJ**, Persson B, Höög J-O. Computational studies of human class V alcohol dehydrogenase — the odd sibling. *BMC Biochem.* 2016;17(1):16.
- IV **Östberg LJ**, Höög J-O, Persson B. Computational analysis of human medium-chain dehydrogenase/reductase revealing substrate- and coenzyme-binding characteristics. Manuscript.

## PUBLICATIONS NOT INCLUDED IN THIS THESIS:

- (A) Hellgren M, Carlsson J, **Östberg LJ**, Staab CA, Persson B, Höög J-O. Enrichment of ligands with molecular dockings and subsequent characterization for human alcohol dehydrogenase 3. *Cell. Mol. Life. Sci.* 2010;67(17):3005–3015.
- (B) Höög J-O, **Östberg LJ**. Mammalian alcohol dehydrogenases—a comparative investigation at gene and protein levels. *Chem Biol Interact.* 2011;191(1–3):2–7.
- (C) Landreh M, **Östberg LJ**, Pettersson TM, Jörnvall H. Transthyretin microheterogeneity and molecular interactions: implications for amyloid formation. *Biomol Conc.* 2014;5(3):257–264.
- (D) Landreh M, **Östberg LJ**, Jörnvall H. A subdivided molecular architecture with separate features and stepwise emergence among proinsulin C-peptides. *Biochem Biophys Res Comm.* 2014;450(4):1433–1438.
- (E) Jörnvall H, Landreh M, **Östberg LJ**. Alcohol dehydrogenase, SDR and MDR structural stages, present update and altered era. *Chem Biol Interact.* 2015;234:75–79
- (F) Barbaro M, Soardi FC, **Östberg LJ**, Persson B, De Mello MP, Wedell A, Lajic S. In vitro functional studies of rare CYP21A2 mutations and establishment of an activity gradient for nonclassic mutations improve phenotype predictions in congenital adrenal hyperplasia. *Clinical endocrinology* 2015;82(1):37–44.
- (G) de Paula Michelatto D, Karlsson L, Lusa AL, Silva CD, **Östberg LJ**, Persson B, Guerra-Júnior G, de Lemos-Marini SH, Barbaro M, de Mello MP, Lajic S. Functional and Structural Consequences of Nine CYP21A2 Mutations Ranging from Very Mild to Severe Effects. *Int J Endocrinol.* 2016;2016:4209670.

# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Bioinformatics . . . . .	1
1.2	Medium-chain Dehydrogenase/Reductase . . . . .	2
1.2.1	Structure . . . . .	3
1.2.2	Function . . . . .	4
1.2.3	Alcohol Dehydrogenase . . . . .	4
<b>2</b>	<b>Methods</b>	<b>7</b>
2.1	In Vitro Studies . . . . .	7
2.2	Databases . . . . .	7
2.3	Sequences . . . . .	8
2.3.1	Comparing Sequences . . . . .	8
2.3.2	Finding Homologs in Sequence Databases . . . . .	10
2.3.3	Evolutionary Relationships . . . . .	12
2.3.4	Conservation . . . . .	13
2.4	Structure . . . . .	13
2.4.1	Predicting Protein Structure . . . . .	13
2.4.2	Studying the Dynamics of a Protein . . . . .	15
<b>3</b>	<b>Present Investigations</b>	<b>17</b>
3.1	Aim . . . . .	17
3.2	The Mammalian Alcohol Dehydrogenases . . . . .	17
3.3	Class V Alcohol Dehydrogenase . . . . .	18
3.3.1	In Vitro . . . . .	19
3.3.2	In Silico . . . . .	19
3.4	Studying the MDR Binding Pocket . . . . .	20
3.5	Summary . . . . .	22
<b>4</b>	<b>General Discussion</b>	<b>25</b>
<b>5</b>	<b>Acknowledgements</b>	<b>29</b>
<b>6</b>	<b>References</b>	<b>31</b>

# LIST OF ABBREVIATIONS

**ADH** Alcohol dehydrogenase

**BLAST** Basic Local Alignment Search Tool

**GFP** Green fluorescent protein

**GST** Glutathione-S-transferase

**HMGS** Hydroxymethylglutathione

**HMM** Hidden Markov model

**MD** Molecular dynamics

**MDR** Medium-chain dehydrogenase/reductase

**MECR** Enoyl-[acyl-carrier-protein] reductase, mitochondrial

**MSA** Multiple sequence alignment

**PDB** Protein Data Bank

**PIR** Protein Information Resource

**PSI-BLAST** Position-specific iterated BLAST

**PSSM** Position-specific scoring matrix

**PTGR** Prostaglandin reductase

**QOR** Quinone oxidoreductase

**QORX** Quinone oxidoreductase PIG3

**RT4I1** Reticulon-4-interacting protein 1, mitochondrial

**SDR** Short-chain dehydrogenase/reductase

**SORD** Sorbitol dehydrogenase

**VAT1** Synaptic vesicle membrane protein VAT-1 homolog

**VAT1L** Synaptic vesicle membrane protein VAT-1 homolog-like



# 1 INTRODUCTION

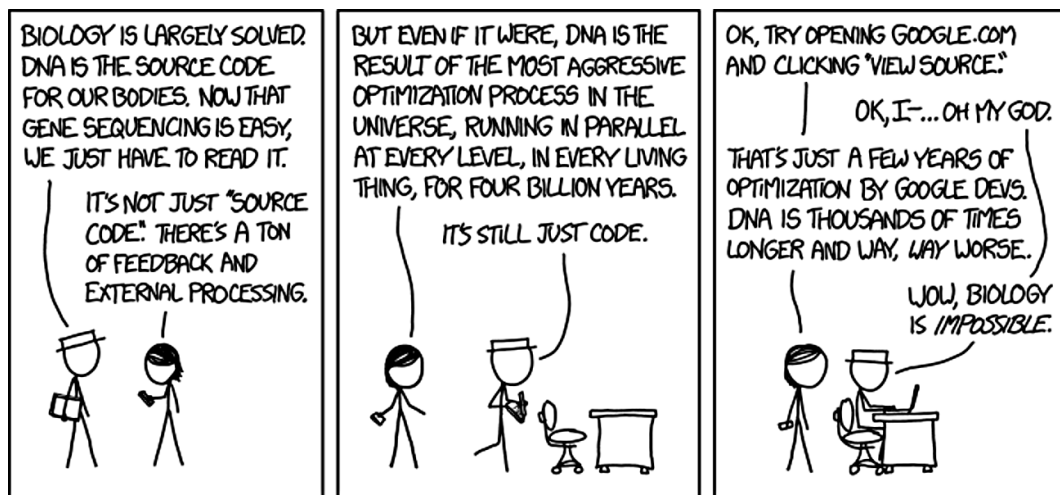


Figure 1.1: xkcd: DNA<sup>1</sup> / CC BY-NC 2.5

As defined by the central dogma of molecular biology<sup>2</sup>, DNA is transcribed into RNA, which in turn is translated into proteins. This means that DNA is used to store the genetic information in the chromosomes. The DNA can be transcribed to multiple copies of RNA as a means to control the expression, and finally each mRNA can be translated into multiple copies of the same protein. The various functions that are needed within an organism are performed by the proteins (and to a minor part, the RNA). As such, the understanding of the function of each protein is essential to fully understand the molecular function of an organism.

## 1.1 BIOINFORMATICS

Traditionally, biomedical processes have been studied using biochemical and biophysical methods. As the knowledge in the field of biochemistry increased, the field of bioinformatics appeared in the 1970s with the goal of making theoretical models that would fit biological data, intended to follow the concept of “theoretical biology”<sup>3</sup>.

Over time, the biochemical methodologies improved, generating more and more data, especially with the advent of high-throughput methods such as next-generation sequencing. In 2003, the sequencing of the human genome was a major milestone<sup>4</sup>, but in 2017, the sequencing of human genomes is done routinely. Similar improvements can be seen in other biochemical fields, including mass spectrometry, structural biology, expression analysis, and many others.

As an effect of the greater amounts of generated data, the definition of the field of bioinformatics transformed, from the original definition of making theoretical models, to being more about analysing biological data. The field of bioinformatics today covers many areas,

including DNA and RNA sequence analysis, protein sequence analysis, phylogeny, structural biology, and systems biology. It also includes the prediction and simulation of different biomedical properties, e.g. calculating structural models of proteins.

The fields of biochemistry and bioinformatics now work in unison, the biochemical methods generating data, which can be evaluated and used for further analysis by the bioinformatics methods.

Along with the generation of scientific data, there is also the need to store the data and to make it available to other researchers.

One of the earliest “systems” for the publication of sequence data was Dayhoff’s Atlas of Protein Sequence and Structure, which in 1969 reported all (initially 65) protein sequences that were known at the time<sup>5</sup>. It was printed as multiple volumes, and led to the development of the Protein Information Resource<sup>6</sup> (PIR), the first protein sequence database accessible by remote computers.

After PIR came more sequence databases. Originally the development efforts were done independently, causing USA, Europe, and Japan to have their own sequence databases. Over time, they started collaborating, and while each site still runs their own databases, the sequence data is now being synchronised. Though the sequences are shared, each site still maintain their own systems with unique user interfaces.

## 1.2 MEDIUM-CHAIN DEHYDROGENASE/REDUCTASE

The medium-chain dehydrogenase/reductase (MDR) superfamily is a protein family with a large number of member proteins, both from eukaryotes and prokaryotes. Due to the large number of members, it is often referred to as a superfamily, its members in turn forming protein families of their own<sup>7</sup>.

The first characterised MDR member was alcohol dehydrogenase, which sequence from horse liver class I was determined in 1970<sup>8</sup>. After the initial member was found, more members were discovered, and the name of the of the superfamily, MDR, was termed to separate the MDRs from another protein family with similar function, the short-chain dehydrogenase/reductase (SDR) superfamily<sup>9</sup>. The MDRs generally have ~350 amino acid residues per subunit (though there are exceptions) while the SDRs generally have ~250 amino acid residues<sup>9</sup>.

The MDR superfamily has considerable multiplicity, with 18 human family members, including the alcohol dehydrogenases (ADH; EC 1.1.1.1), prostaglandin reductases (PTGR; 1.3.1.48), quinone oxidoreductases (QOR; EC 1.3.1.27), sorbitol dehydrogenases (SORD; EC 1.1.1.14), as well as multiple other enzymes (Table 1.1). The number of proteins increases as more species are analysed. Multiple attempts to classify the family have been made. One such separated the 15,000 sequences known in 2010 into 86 unique clusters with at least 20 protein sequences each<sup>10</sup>, and a large number of small clusters with fewer member that were not evaluated at the time. The total number of unique families has been estimated to ~500 with ~30% sequence identify between the families<sup>11</sup>.

There are three domain definitions of the MDR proteins in Pfam<sup>12</sup>, each covering different members and parts of the proteins; ADH\_N (Alcohol dehydrogenase GroES-like domain; PF08240), ADH\_N\_2 (N-terminal domain of oxidoreductase; PF16884), and ADH\_zinc\_N (Zinc-binding dehydrogenase; PF00107). Together, the definitions identify, as of February 2017, more than 170,000 MDR proteins in the UniProt sequence database<sup>IV</sup>, a number that increases with each new release of the database.

**Table 1.1:** Human MDR members, using UniProtKB/Swiss-Prot annotations.

Protein name	Gene	UniProt	Structure <sup>1</sup>	Coenzyme
Alcohol dehydrogenase 1A	ADH1A	P07327	1U3T (2)	NAD(H)
Alcohol dehydrogenase 1B	ADH1B	P00325	1U3V (9)	NAD(H)
Alcohol dehydrogenase 1C	ADH1C	P00326	1U3W (3)	NAD(H)
Alcohol dehydrogenase 4	ADH4	P08319	3COS (1)	NAD(H)
Alcohol dehydrogenase 6	ADH6	P28332	Unknown	Unknown
Alcohol dehydrogenase class-3	ADH5	P11766	2FZW (9)	NAD(H)
Alcohol dehydrogenase class 4	ADH7	P40394	1D1T (3)	NAD(H)
mu/sigma chain				
Enoyl-[acyl-carrier-protein] reductase, mitochondrial	MECR	Q9BV79	1ZSY (2)	NADP(H)
Fatty acid synthase	FASN	P49327	3TJM (12)	NADP(H)
Prostaglandin reductase 1	PTGR1	Q14914	2Y05 (2)	NADP(H)
Prostaglandin reductase 2	PTGR2	Q8N8N7	2ZB4 (6)	NADP(H)
Prostaglandin reductase 3	PTGR3	Q8N4Q0	2C0C (4)	NADP(H)
Quinone oxidoreductase	CRYZ	Q08257	1YB5 (1)	NADP(H)
Quinone oxidoreductase PIG3	TP53I3	Q53FA7	2J8Z (2)	NADP(H)
Reticulon-4-interacting protein 1, mitochondrial	RTN4IP1	Q8WWV3	2VN8 (1)	NADP(H)
Sorbitol dehydrogenase	SORD	Q00796	1PL8 (3)	NAD(H)
Synaptic vesicle membrane protein VAT-1 homolog	VAT1	Q99536	Unknown	Unknown
Synaptic vesicle membrane protein VAT-1 homolog-like	VAT1L	Q9HCJ6	4A27 (1)	NADP(H)

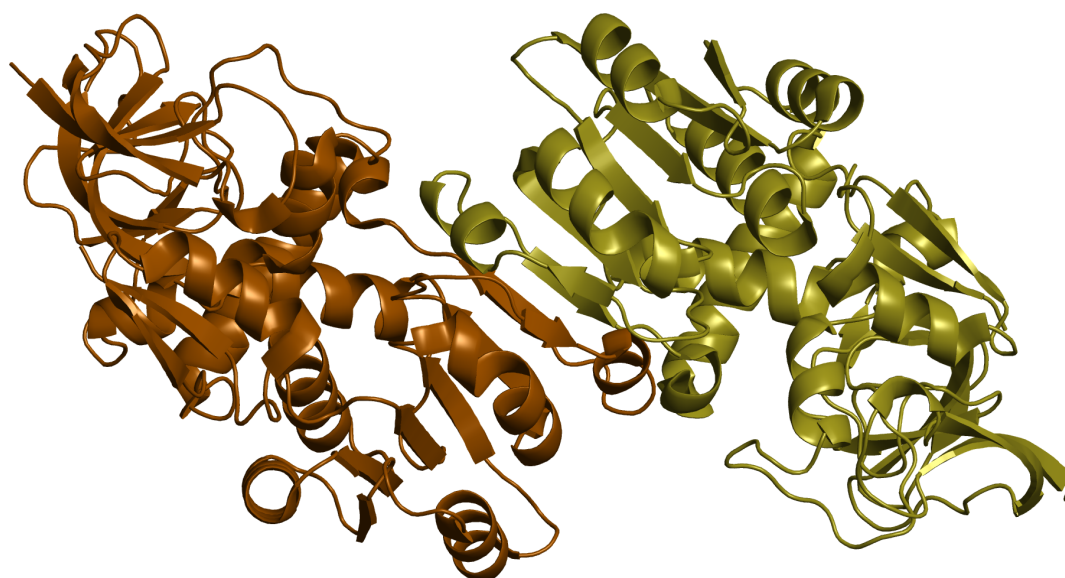
<sup>1</sup>PDB code of one structure (total number of structures in PDB)

## 1.2.1 Structure

The structure of an MDR protein is formed by two domains; one C-terminal coenzyme (NAD(H)/NADP(H)) binding domain with the common Rossmann fold<sup>13</sup>; typically six  $\beta$ -strands forming a  $\beta$ -sheet with  $\alpha$ -helices above and below, and one substrate-binding domain consisting of antiparallel  $\beta$ -strands with  $\alpha$ -helices at the surface, similar to the GroES fold<sup>14</sup>, the binding pocket being located inside the cleft between the two domains.

MDRs often form homodimers, but both monomeric (e.g. mannitol dehydrogenase<sup>15</sup>) and tetrameric (e.g. yeast alcohol dehydrogenase I<sup>16</sup>) members are known to exist.

Currently, for 61 of the MDR proteins listed in UniProtKB/Swiss-Prot, there is an X-ray crystal structure available, and the number is increased to 117 if the proteins in UniProtKB/TrEMBL are also included. Of these 117 proteins, 16 are the human forms, meaning that the structures are still unknown for only two human MDR proteins. It also means that a majority of all known MDRs does not have known structure, though many are expected to be homologous to the known structures.



**Figure 1.2:** Typical dimeric MDR structure: human QOR (PDB id: 1YB5).

## 1.2.2 Function

As there are a multitude of different MDRs, there is also a large variety of functions. The characterised MDR members generally function as enzymes catalysing the reaction  $\text{Alcohol} + \text{NAD(P)}^+ \rightleftharpoons \text{Aldehyde/ketone} + \text{NAD(P)H}$ , where the enzymes using  $\text{NAD}^+$  typically function as dehydrogenases, while the ones using  $\text{NADPH}$  function as reductases<sup>11</sup>. Though the function of many MDR members are known, there are many more with unknown function.

The MDRs are homologs, though their sequences show a great level of variation, but the proteins all share similar folds. However, the binding pockets can vary quite a bit. The mammalian alcohol and sorbitol dehydrogenases use  $\text{Zn}^{2+}$  ions to help with the coordination of the substrates<sup>17</sup>, while many other MDRs use Tyr<sup>18</sup>. There are different types of binding pockets among the proteins using Tyr, each containing the Tyr at different positions, and some including a second Tyr<sup>IV</sup>.

The substrate specificities vary greatly among the MDRs, from ethanol<sup>19</sup> to benzoquinone<sup>19</sup>, prostaglandins<sup>20</sup>, fatty acids<sup>21</sup>, and trans-2-hexenoyl-CoA<sup>22</sup>.

## 1.2.3 Alcohol Dehydrogenase

Alcohol dehydrogenase is the largest of the MDR families. In the 2010 classification of MDRs<sup>10</sup>, there were more than 2200 sequences in the main cluster, as well as hundreds in smaller clusters. As the total number of MDR sequences since then has grown by an order of magnitude, the number of ADH sequences are expected to have increased by a considerable amount as well.

There are multiple isoforms, classes, present in different species. In mammals, there are six classes, class I–VI (ADH1–6) with  $\sim 60\%$  sequence identity between the classes. The classes is a common source of confusion, as the numbering is based on the order of discovery, and the proteins and the genes do not necessarily have the same number. This thesis uses the classification recommended in 1999<sup>23</sup>, where the class I ADH is annotated as ADH1, class II as ADH2, and so forth. In humans, there are seven ADH genes, corresponding to ADH1–5<sup>19,24</sup>, with three isoforms of ADH1<sup>25</sup> (Table 1.2).

The human ADHs are mainly expressed in the liver, but may be found in varying amounts in other tissues as well. The exceptions are ADH3 which can be found in most tissues, and ADH4 which is not found in the liver, but in the stomach and the intestines<sup>26</sup>.

The substrate specificities of the classes vary. The ADH1 isoforms are the traditional enzymes involved in the breakdown of ethanol to acetaldehyde, but also have activity with e.g. retinol, hydroxysteroids, and hydroxytryptophol. ADH2 has a high activity with ethanol, but with a high  $K_m$ <sup>19</sup>. The preferred substrates are retinol<sup>27</sup> and benzoquinone<sup>28</sup>. ADH3 acts especially on S-nitrosoglutathione and S-hydroxymethylglutathione, but its substrates also include  $\omega$ -hydroxy fatty acids, and, with a very high  $K_m$ , ethanol<sup>29</sup>. ADH4 has a preference for ethanol and retinoids<sup>30</sup>. As for ADH5 and ADH6, the function is currently unknown, and the proteins have never been isolated in their native forms.

**Table 1.2:** Nomenclature for the human alcohol dehydrogenases

The Class System		Suggested System <sup>1</sup> Protein and Gene	UniProt	Gene Nomenclature	
Protein				New	Old
Class I	$\alpha$ -subunit	ADH1A	P07327	ADH1A	ADH1
Class I	$\beta$ -subunit	ADH1B	P00325	ADH1B	ADH2
Class I	$\gamma$ -subunit	ADH1C	P00326	ADH1C	ADH3
Class II	$\pi$ -subunit	ADH2	P08319	ADH4	ADH4
Class III	$\chi$ -subunit	ADH3	P11766	ADH5	ADH5
Class IV	$\mu/\sigma$ -subunit	ADH4	P40394	ADH7	ADH7
Class V		ADH5	P28332	ADH6	ADH6

<sup>1</sup>As proposed by Duester et al<sup>23</sup>

Based on the theory of enzymogenesis, where the ancestral form of an enzyme tends to have a higher level of conservation than any duplicates of the gene<sup>31</sup>, ADH3 is considered the ancestral form of the mammalian ADHs, with an average sequence identity of 93.4% among the known sequences, compared to 83.6% of ADH1 and 77.8% of ADH5<sup>II</sup>.

In humans, the genes encoding the ADHs are located on chromosome 4 (4q23), and the gene order is ADH3-ADH2-ADH5-ADH1A-ADH1B-ADH1C-ADH4<sup>32</sup>. The gene order is the same in other mammals, and in the species that have ADH6, it is present between ADH5 and ADH1.

Mammalian ADHs have the typical MDR structure, but has an insertion of 50 residues (or other MDRs have a deletion, as the MDR structure was originally defined by ADH) at position 86 (PTGR1 numbering), forming a loop surrounding a  $Zn^{2+}$  ion. The zinc ion is kept in place by four deprotonated cysteines; Cys97, Cys100, Cys103, and Cys111 (human ADH1C numbering, primarily from horse class I ADH), and is called the structural zinc for its function as a stabiliser of the mammalian ADH structures. The mammalian ADHs have another zinc ion located at the active site, the catalytic zinc, where it supports the catalytic reaction. That zinc ion is stabilised mainly by two cysteines, Cys46 and Cys174, as well as by His67<sup>33</sup> (human ADH1C numbering, primarily from horse class I ADH). The mammalian ADHs form homodimers, and in rare cases heterodimers, while some ADHs from other species, e.g. yeast ADH1, may form tetramers<sup>16</sup>.



## 2 METHODS

### 2.1 IN VITRO STUDIES

A common approach to study a protein in vitro is to express it in sufficient quantities and then isolate it. This can be done by inserting the target gene into an expression vector and then getting the vector into the cell type of choice. Upon insertion, the genes on the vector will be expressed by the normal systems inside the cell.

In order to extract the protein, the cells are commonly lysed by e.g. sonication, and the proteins are separated from the rest of the cell lysate by using e.g. gel columns that separate by size or affinity to certain molecules.

The isolated proteins can be studied using conventional assays.

Gene expression in tissue samples can be studied using Northern blot<sup>34</sup>. The cells of interest are lysed, followed by isolation of the RNA from the rest of the lysate. The RNA is added to a gel and separated by size, followed by transfer to a nylon membrane. Labelled DNA or RNA probes matching the sequence of interest are then hybridised against the RNA. Finally the labelling of the probe is used to detect the wanted RNAs.

### 2.2 DATABASES

In the case of nucleotide sequences (DNA/RNA), the major databases available today are the European Nucleotide Archive (ENA)<sup>35</sup>, Genbank<sup>36</sup>, and the DNA Data Bank of Japan (DDBJ)<sup>37</sup>. There are also multiple other databases focusing on specific species as well as genomes, e.g. Ensembl<sup>38</sup>, where the main database is focused on mainly vertebrate genomes, but there are also sister databases focusing on plants, protists, fungi, other metazoa, and bacteria. Genbank release 218 (Feb 2017) contains 199,341,377 nucleotide sequences.

In the case of proteins, UniProt<sup>39</sup> and NCBI Protein<sup>36</sup> are the major databases. UniProt has two parts; UniProtKB/Swiss-Prot, originally made by the Swiss Institute of Bioinformatics, which has protein sequence entries that have been manually reviewed and annotated, and UniProtKB/TrEMBL, which is annotated using automated methods. The UniProt database contains, as of release 2017\_03, 553,941 entries in UniProtKB/Swiss-Prot and 80,204,459 entries in UniProtKB/TrEMBL.

As the size of the databases keep on increasing, the need of automated approaches becomes more and more pronounced. In most cases, automated annotations by e.g. sequence similarity work well, but the methodology may sometimes cause incorrect classification of enzymes. One such case may be when the definitions do not contain information about new classifications. One example is the case of class V and class VI ADH. The automated annotations for mammalian ADHs expect only five classes, and proteins that in reality belong to class VI will be mistaken for members of other classes (see section 3.2).

There are also a lot of specialised databases in regard to sequences. Pfam<sup>12</sup> contains definitions of protein families and domains, and maintains information about e.g. sequences and

structures that have been mapped to the definitions. Interpro<sup>40</sup> integrates the definitions provided by Pfam, as well as other domain databases, allowing automated protein classification and prediction of domains.

Apart from databases for sequences and their annotations, there are also databases for many other types of biological data. Some examples include the Protein Data Bank (PDB)<sup>41</sup> containing experimental protein structures, the Peptide Atlas<sup>42</sup> providing mass spectrometry data mapped to proteins, and the Human Protein Atlas<sup>43</sup> providing information about where human proteins are expressed. There are also databases focused on the properties of small molecules, e.g. PubChem<sup>44</sup>.

## 2.3 SEQUENCES

### 2.3.1 Comparing Sequences

In order to find out if two proteins are e.g. similar or contain similar domains, it would be useful to compare their amino acid residue sequences. The task of comparing two sequences and lining them up (aligning them) to match each other is commonly referred to as sequence alignments. In the case of very closely related proteins with identical length, the proteins can just be aligned residue by residue, but if the sequence lengths vary due to e.g. insertions or deletions, the differences in the length must be considered. This is usually done by inserting empty positions, gaps, in the positions that are missing.

The simplest form of a sequence alignment is a pairwise alignment, where two sequences, nucleotides or amino acid residues, are aligned. If more sequences are aligned, the alignments are commonly referred to as multiple sequence alignments (MSA).

The core problem of performing a sequence alignment is evaluating which units (nucleotides or amino acid residues) that should be aligned, and where gaps should be inserted. A very common solution for this is to use a scoring matrix defining the matches between the different units, i.e. defining that a match between e.g. an A and A would give the score 5, while a match between A and C would instead give the score -1. There is a multitude of different scoring matrices available, often optimised for certain situations such as a specific species. In the case of nucleotides, the scoring may be based on e.g. GC content. In the case of proteins, the scoring matrices tend to be more complex, where the BLOSUM matrices<sup>45</sup>, especially BLOSUM62, are the most common. The scores in the BLOSUM matrices are derived from MSAs of sequences that have been clustered at a certain level of sequence identity, e.g. 62% in the case of BLOSUM62. The alignments have then been analysed in regard to how common the different amino acid residue substitutions are, and the final scores calculated.

In the case of gaps, there are two main approaches; linear and affine gap penalties (penalty as the score should always be worse than single matches). Linear gap penalties give the same penalty wherever a gap is added into the alignment. The affine gap penalties are divided into two parts: an insertion penalty and an extension penalty, where the insertion penalty is higher than the extension one. The theory is that it's more probable to have many gaps in the same stretch, rather than split up throughout the sequence.

Finally, there are two main types of sequence alignments: global and local ones. The global alignments align all sequences in their entirety, while the local alignments focus on high-scoring regions, e.g. region with a high similarity due to sharing the same domain.

Testing all possible alignments to find the best (optimal) one would be a very computationally expensive approach, as even two 100 amino acid residue protein sequences can be aligned in  $\sim 10^{75}$  different ways (with gaps). Thus fast algorithms are needed.



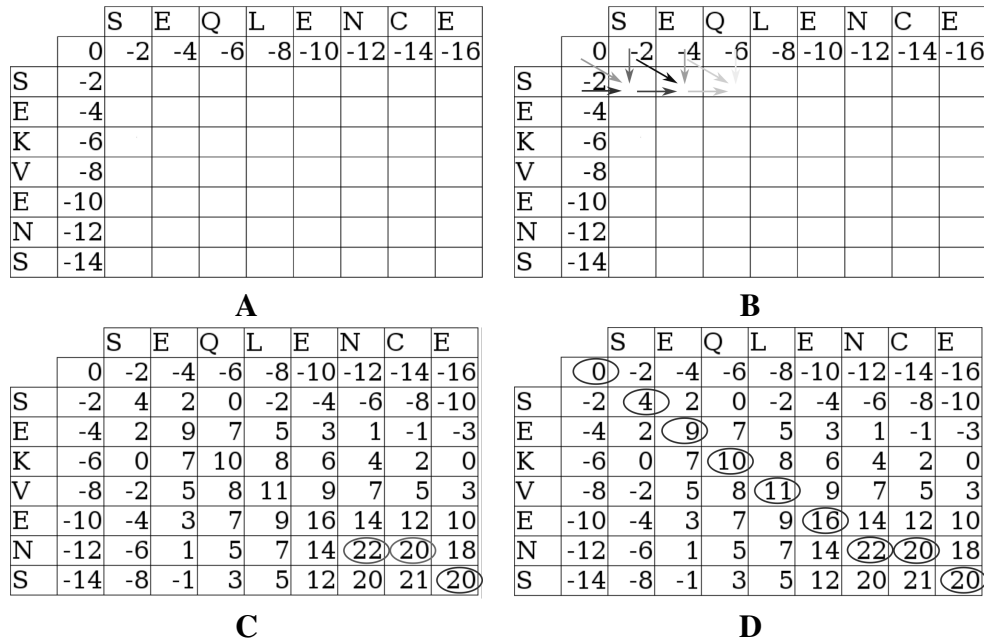
## Pairwise Alignments

In the case of pairwise alignments, the algorithm for identifying the alignment with the optimal score given a scoring matrix and gap penalty can be solved by dynamic programming. Global pairwise alignments are generated by the Needleman-Wunsch dynamic programming algorithm<sup>46</sup>.

The dynamic programming algorithm is based on the calculation of the optimal path for each field in a matrix. The initial step is to generate an empty matrix, where the fields ( $F$ ) correspond to the sequences. The starting point  $F(0, 0)$  is set to 0. The score of the rest of the fields are then depending on the neighbours directly above (gap), to the left (gap), and to the upper left (match). The score in a field is calculated according to equation 2.1, where  $s$  is the value from the scoring matrix for a match between the residues. As  $F(i, 0)$  and  $F(0, j)$  can only be reached by gaps, these fields are calculated immediately (Figure 2.1:A). The rest of the score is then calculated recursively for each of the fields, from  $F(1, 1)$  to the lower right corner (Figure 2.1:C).

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases} \quad (2.1)$$

When all fields are calculated, a traceback operation is performed. Starting in the lower right corner (as the goal is a global alignment), the path to get the score in that field is traced back (Figure 2.1:C). The traceback finishes when the  $F(0, 0)$  is reached. The alignment corresponds to the path of the traceback (from  $F(0, 0)$ , a diagonal being a match, a movement right indicating a gap in the left sequence, and a movement down implying a gap in the top sequence).



**Figure 2.1:** Dynamic programming algorithm for the sequences SEKVEN and SEQUENCE with BLOSUM62 and gap penalty:  $-2$ . A: Adding default gaps. B: Calculation of the matrix; match/gap/gap for each field. C: Matrix fully calculated, starting traceback. D: Calculations done. The final alignment is SEQUENCE vs SEKVEN-S.

The algorithm for pairwise local alignments is very similar to the one for global alignment<sup>47</sup>. The basic algorithm is the same, but the scoring step has another condition added

(equation 2.2) and the traceback starts in the field with the highest score and finished when it reaches a field with 0.

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \\ 0 \end{cases} \quad (2.2)$$

The process of aligning two sequences, called pairwise alignment, is generally considered solved by dynamic programming<sup>46,47</sup>, resulting in the optimal alignment based on the given parameters, i.e. scoring matrix and gap score.

There are other algorithms for performing pairwise alignments, e.g. FOGSAA<sup>48</sup>, which are much faster than the original dynamic programming approaches, but are not guaranteed to always provide the optimal alignment. The original methods thus still have a place to fill.

## Multiple Sequence Alignments

It is often of interest to align more than two sequences in order to obtain more information about the sequence properties, as well as to study the similarities within e.g. a protein family.

Sadly, dynamic programming cannot be efficiently extended to aligning more than two sequences, as the number of required computations increases exponentially with the number of sequences. Thus, a number of alternative approaches have been developed to make MSAs, sacrificing the promise of providing optimal alignments for speed. Many of the current methods start by comparing the sequences pairwise, generating a guide tree showing which sequence have the highest level of similarity. The alignment is then generated by adding the sequences to the alignment in the order given by the guide tree, a process called progressive alignment. The accuracy of the alignment can be improved by e.g. iterative realignment processes during the “alignment” merging. Modern methods often include approaches as well, including e.g. fast Fourier transformations and HMMs.

There are multiple MSA methods available, and the choice of method is often a balance between accuracy and speed. The accuracy is also dependent on the properties of the sequences, e.g. the amount of insertions, different methods providing better alignment for different types of sequences. The accuracy of the methods can be evaluated by e.g. manually curated alignments such as BALiBASE<sup>49</sup>, but simulated sequences are also common in order to get larger sequence sets with controlled properties. Based on a benchmarking of multiple modern MSA methods<sup>50</sup>, the most accurate methods include ProbCons<sup>51</sup> and the L-INS-i approach of MAFFT<sup>52,53</sup>, while e.g. Kalign<sup>54,55</sup> and the FFT-NS-2 approach of MAFFT are very fast (but less accurate). Depending on the types of sequences, other methods such as MulAlin<sup>56</sup> and SATe<sup>57,58</sup> are also good choices. Clustal Omega<sup>59</sup> is often a good trade-off between speed and accuracy.

### 2.3.2 Finding Homologs in Sequence Databases

When working with proteins, a very common objective is to find homologs to the current protein in the major sequence databases, e.g. UniProt and NCBI. There have been multiple methods developed for this purpose, both general and specialised ones.

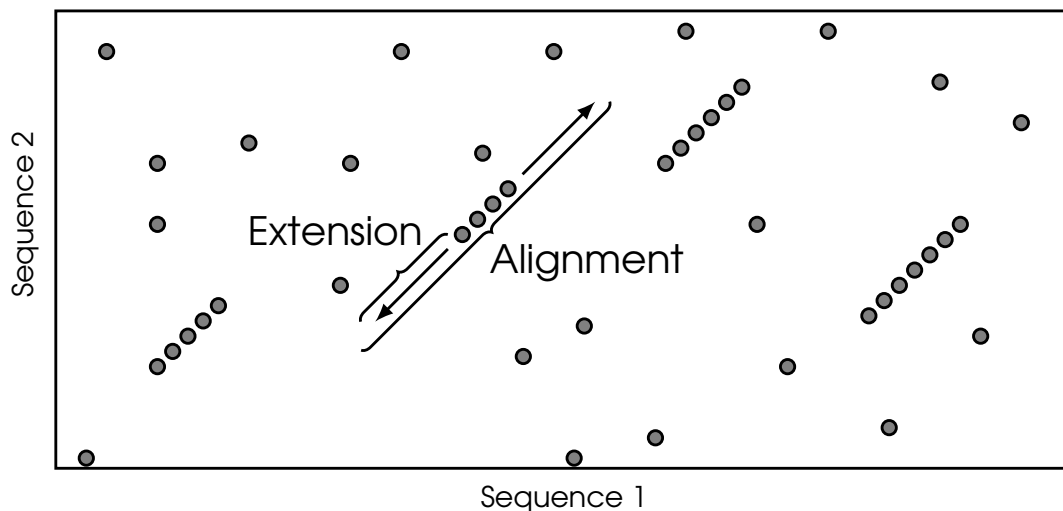
## BLAST

The de facto standard for performing sequence searches in databases is the Basic Local Alignment Search Tool (BLAST)<sup>60,61</sup> suite of programs, covering searches for both nu-

cleotide (blastn) and protein (blastp) sequences, and even automatic translation of nucleotide sequences to protein sequences (blastx) and vice versa (tblastn).

The original BLAST algorithm was published in 1990<sup>62</sup> and is based on the concept of words (short regions with similar, not necessarily identical, residues), in contrast to the main sequence search software that existed at the time, FASTA<sup>63</sup>, which was searching for kmers of identical residues. This allowed BLAST to identify more distant homologs than was previously possible.

The algorithm starts by identifying words that would score above a certain threshold (using e.g. BLOSUM62), followed by matching of the words towards a sequence from a database. Regions with multiple words are extended into local alignments using dynamic programming. The local alignments with scores above a certain threshold are then reported as hits. The quality of a BLAST hit is commonly presented as an E value representing the probability of discovering a sequence with the current score in a database of the current size, but for e.g. short hits, other evaluation methodologies are commonly used, e.g. the sequence identity.



**Figure 2.2:** The BLAST algorithm. The sequences are matched against high-scoring words (gray), and regions with multiple words are used as seeds for a local alignment that is extended until a defined threshold. Local alignment above the score threshold are then reported as results. *Modified from image by Per Unneberg*

The standard BLAST method performs well for close homologs, but lacks sensitivity in regard to distant homologs. For this purpose, an iterative version of BLAST was developed; position-specific iterative BLAST (PSI-BLAST)<sup>60</sup>. PSI-BLAST starts by performing a normal BLAST search using e.g. BLOSUM62 as scoring matrix, but then the alignments of the results from the first search (iteration) are merged and a position-specific scoring matrix (PSSM) is calculated, keeping the data for the positions in the original query sequence. Another iteration is then started, but the alignments are scored using the PSSM instead of the original scoring matrix. The search can then be iterated, generating a new PSSM with each iteration, until the search has converged (no new sequences are found).

Further specialised versions of BLAST have also been developed, including e.g. DELTA-BLAST<sup>64</sup>, which adds information from domain databases to the search to make it more accurate and sensitive for distant homologs.

## HMMER

A protein or nucleotide sequence can be described as a hidden Markov model (HMM). By adding states corresponding to the residues/nucleotides at each position, as well as for insertions and deletions, the HMM can then be used to evaluate the similarity of other sequences to the original sequence. If an MSA containing sequences from e.g. the same proteins family is used to generate the HMM, the HMM can thus be used to evaluate whether other sequences belong to the same family. A common program for this purpose is HMMER<sup>65</sup>, which also has a special module called JACKHMMER which can be used for the same purpose as BLAST, i.e. to find similar sequences in sequence databases.

Using the HMMs generated by HMMER, it is possible to create automated classifiers for protein families, as have been done for e.g. domains in the database Pfam<sup>12</sup>.

Using HMMs is also useful when the goal is to separate e.g. a protein family into its members. E values and similar scores evaluate the probability of a sequence being a homolog, which should be true for all the members within the family. If an HMM is defined for each of the possible member proteins, then the HMM having the highest probability for a match should be the correct member, as was done for the six classes of ADH in paper II.

### 2.3.3 Evolutionary Relationships

It is often of interest to study the evolutionary relationships between proteins. In its traditional form, phylogeny is used to study the evolutionary relationship between species, but phylogenetic methods, e.g. PhyML<sup>66</sup>, can also be used to study the relationships between protein families.

Phylogenetic trees visualise the evolutionary relationships between species or proteins. All species or sequences, depending on what is compared, are visualised as nodes in a tree graph, while the edges show the relations. There are multiple different types of trees, but many scale the branches according to the “evolutionary distance” between the species or sequences that are compared.

### Evolutionary Pressure

Evolution is an ongoing process, and all nucleotide and protein sequences will change over time to a varying degree. As the function of a gene is most commonly performed by the encoded protein, the preservation of the amino acid sequence is of importance. If the sequence changes in a way that makes the protein less efficient for its purpose, the organism will have an evolutionary disadvantage, and thus there is an evolutionary pressure to maintain the same sequence. The mutations occur at the DNA level, and as many amino acid residues are encoded by multiple different codons, a change may either be non-synonymous or synonymous, i.e. the mutation will or will not change the encoded residue.

If the function of a protein has become less important due to e.g. changes in the environment surrounding the organism, the evolutionary pressure will decrease, as the loss of efficiency for the protein will not cause the same evolutionary disadvantage, and thus more non-synonymous mutations are accepted. As the mutations occur over time, a gene that recently lost its importance may still have the characteristics of a gene. One way of evaluating the level of evolutionary pressure is to compare nucleotide sequences from different species and simulate how they were changing, and then calculating the quotient  $\frac{\text{non-synonymous}}{\text{synonymous}}$ , or in short  $\frac{dN}{dS}$ . If the result is greater than 1, i.e. there are more non-synonymous than synonymous mutations, it strongly implies that the gene is “disappearing” from the genome. As a reference, the mammalian members of the highly conserved histone H1 has a ratio of 0.055,

while the highly variable interleukin-2 has a ratio of 0.556<sup>III</sup>. There are many methods for performing such calculations, a common one being PAML<sup>67</sup>.

### 2.3.4 Conservation

Connected to the concept of evolution is the concept of conservation: what parts of the sequence are conserved by the evolutionary pressure?

The common approach to evaluate the conservation is to collect sequences from multiple species, perform a multiple sequence alignment and evaluate the results. The important residues in a protein are often conserved even in distantly related species, as well as within a protein family.

How distantly related the sequences that are included in the alignment should be is based on what property is of interest. In the case of MDRs, the properties of the binding pocket for NAD(P)H is expected to be similar between the species, and comparing the conserved residues in proteins binding NAD(H) with the ones in proteins binding NADP(H), it is possible to identify residues are responsible for the selection of coenzyme.

The same approach can theoretically be used to identify the active site in an enzyme, but the choice of included sequences is then very important, as evidenced by the results in Paper II. If the sequences do not share the same type of active site, the obtained results will be inaccurate.

Finally, the conserved residues can be mapped to structures to evaluate their localisation; are they part of the active site, important for protein-protein interaction, or maybe have some other functions?

## 2.4 STRUCTURE

While the sequence of a protein is the raw results of the expression of a gene, the way it folds is crucial for its function. Thus, the structure of a protein is more conserved than its sequence. It also means that there is a strong incentive to determine as many protein structures as possible.

The common methodology for determining structures are X-ray crystallography, NMR, and electron microscopy, X-ray crystallography being the most common with nearly 90% of the structures in the PDB. The number of determined structures is growing, and as of March 2017, there are more than 125,000 structures deposited to the PDB<sup>68</sup>.

### 2.4.1 Predicting Protein Structure

The structure has been determined for a large number of proteins, but the structures of a vastly greater number of proteins remain to be determined, and the number of proteins without known structure increases much faster than the number of proteins with known structures.

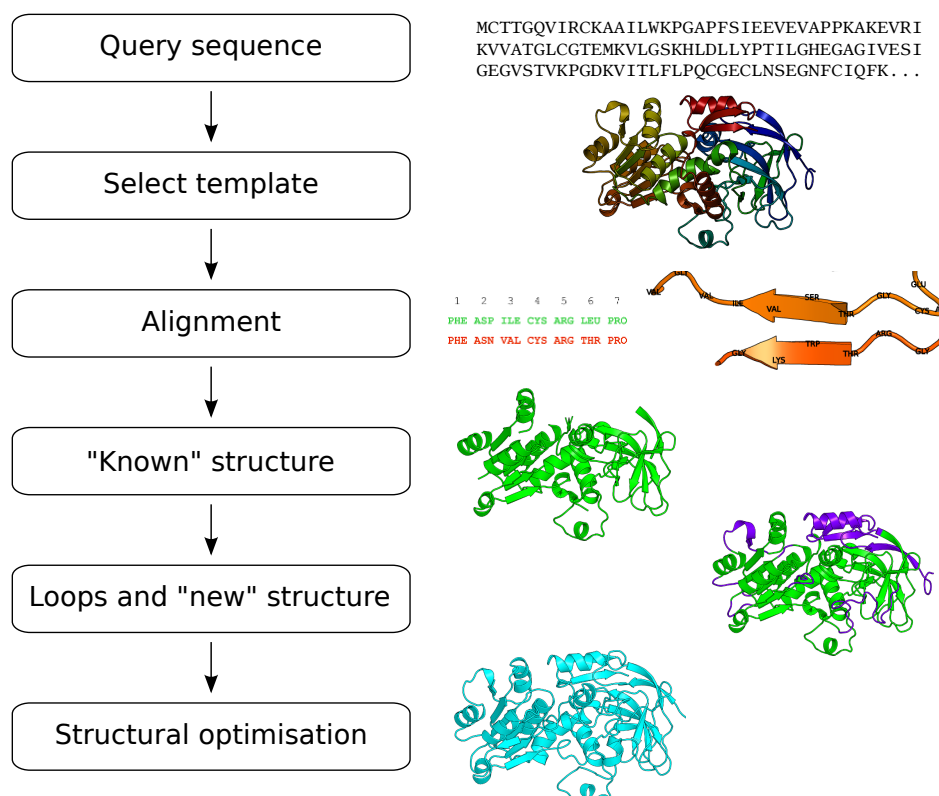
One way of bypassing this problem is to use computational methods to predict the structure based on its sequence. Sadly, the folding of a protein cannot be solved by a simple and straight-forward approach. The protein folding (Levinthal's) paradox<sup>69</sup> states that the number of possible conformations available to a given protein is extremely large, and a brute-force method exploring all conformations of even a small 100-residue protein would take longer to finish than the universe has existed. Thus, an algorithm attempting to predict a protein structure would need to use shortcuts.

There are a multitude of methods developed for ab initio structure predictions with energy-based and fragment-based approaches, using physical principles to predict a probable structure, often requiring considerable computational resources.

Another ab initio approach that has been gaining a lot of traction is to use evolutionary covariation for the predictions, the idea being that residues that mutate together should be located near each other in the protein structure. This approach has been quite successful and there are multiple methods developed, e.g. EVfold<sup>70</sup>. There are also methods that combine multiple different approaches, such as PconsE, which combine the coevolutionary analysis of PconsC<sup>71</sup> with the ab initio-approach used in the Rosetta suit of programs<sup>72</sup>.

There is also another type of protein structure prediction method that attempts to use proteins with known structure as templates for the prediction, called homology modelling (or sometimes template-based modelling).

In its simplest form, a homology modelling method would use the structure of a homolog to the protein of interest and copy the backbone from the parts that have high levels of similarity between the structures. The method would then connect the copied parts by generating the missing regions using e.g. ab initio methods, but with the template structure as a starting point. The generated structure model would then be refined, often using e.g. energy minimisation.



**Figure 2.3:** Schematic representation of a general homology modelling method

As homology modelling uses homologs as templates, the methodology is heavily dependent on the availability of close homologs with known structures. There are also variations of the template identification approach that matches the sequence to known structures, evaluating whether the sequence could form the structure in question, rather than evaluating whether they are homologous.

There are a large amount of homology methods available, including e.g. MODELLER<sup>73</sup>, SWISS-MODEL<sup>74</sup>, and I-TASSER<sup>75</sup>. More methods are constantly being developed, improved by new algorithms, not to mention the increasing amount of possible template struc-

tures in the PDB.

In the case of MDRs, the standard homology modelling methods often provide very good models in regard to the backbone of the protein. However, as many MDRs also bind cofactors, e.g. NAD(H)/NADP(H) and  $\text{Zn}^{2+}$  ions, which are difficult to add after the model refinement steps, the models generated by the standard methods are often suboptimal for MDR enzymes.

In the projects forming this thesis, the most common approach was to use a modified homology modelling approach from the ICM software. First a protein structure model was created using the standard ICM method<sup>76</sup>, whereafter the coenzyme(s) were added before refinement, during which all sidechains were forced into positions that allowed the coenzyme to stay in place before constraints were gradually relaxed and the protein structure model was allowed to find its local energy minimum.

## 2.4.2 Studying the Dynamics of a Protein

When a protein is successfully crystallised, the crystal contains a large amount of individual copies of the protein in the same (or similar) conformations. When the crystal is scanned with X-rays, the refraction patterns can be used to determine the structure. If the individual copies of the protein are all in the exactly same conformation, the refraction pattern will be stronger and the structure can be inferred. If the conformation of a part of the protein varies between the individual protein copies, there will be multiple refraction patterns, and all of them will be weaker, making it more difficult to give an accurate representation of that part of the structure.

As an effect of this, all crystal structures are static, i.e. showing only one conformation of the proteins. As proteins are dynamic, constantly varying the structure to allow e.g. the binding of substrates, a crystal structure cannot represent the full array of different conformations a protein can have. The dynamics of a protein can be studied using NMR, though it mainly works on small proteins. Instead, the dynamics of a protein can be simulated in silico using molecular mechanics.

The two main approaches in molecular mechanics for proteins are Monte Carlo and molecular dynamics (MD) methods. The Monte Carlo methods performs random changes to the system and evaluates their results, generating an ensemble of configurations, giving information about different states for the protein<sup>77</sup>. The MD methods instead simulate the movement of the atoms over short timesteps.

The general properties and equations describing the properties and interactions between atoms are often referred to as the force field. Any calculations on a molecular system would preferably use quantum mechanics, but as those calculations are very computationally intensive, they are infeasible for large systems such as proteins, especially when the behaviour over larger time scales are simulated. Instead the methods uses approximations that are considered to be accurate enough, but may lack in their exact representation of e.g. hydrogens. Thus, molecular dynamics uses Newtonian (classical) mechanics and approximate representations of e.g. van der Waal interactions to calculate how the system behaves over time.

As proteins are commonly not performing their functions in a vacuum, it is a good idea to simulate e.g. water around the protein molecule, either explicitly (with added water molecules) or implicitly (using general equations to describe the interactions of the protein with a general water environment). The addition of water molecules adds a lot of new atoms to the calculations, increasing the amount of required calculations when compared to implicit water, but may generate more accurate results by e.g. explicitly simulating water molecules that are interacting with the protein.

GROMACS<sup>78</sup> is an open source implementation of a molecular mechanics system, performing energy minimisation (finding local energy minima) and molecular dynamics, that

was originally developed at the Groningen University in 1991. It is able to use different force fields and is highly optimised for speed by using optimised CPU instructions as well as supporting heterogenous systems (CPU + GPU) for a further performance boost<sup>78</sup>.



## 3 PRESENT INVESTIGATIONS

### 3.1 AIM

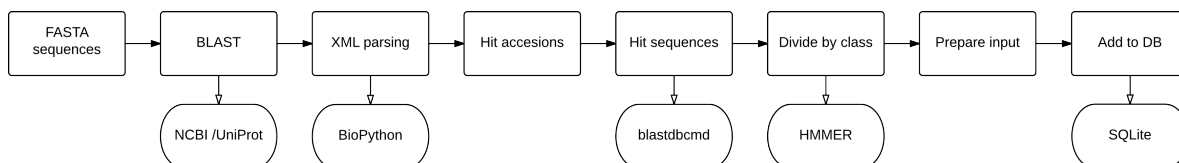
- To identify mammalian alcohol dehydrogenases in the databases and to investigate that their annotations match their real classification (class I–VI), including analysis of the identified members and their relationships (paper II; section 3.2).
- To study ADH5, first attempting to isolate and characterise it, followed by computational analysis of the structure of human ADH5 as well as the sequences of ADH5 from other mammals (paper I and III; section 3.3).
- To define important residues in the binding pockets of MDRs, focusing on nine different MDRs that are present in humans and for which only little is known (paper IV; section 3.4).

### 3.2 THE MAMMALIAN ALCOHOL DEHYDROGENASES

With the constant inflow of new sequences to the UniProt and NCBI sequence databases, there is also a constant inflow of new mammalian ADH sequences. Most of the sequences are correctly annotated, but there are also multiple ones that are not annotated at all (“uncharacterized protein”), or in some cases annotated as belonging to the incorrect class.

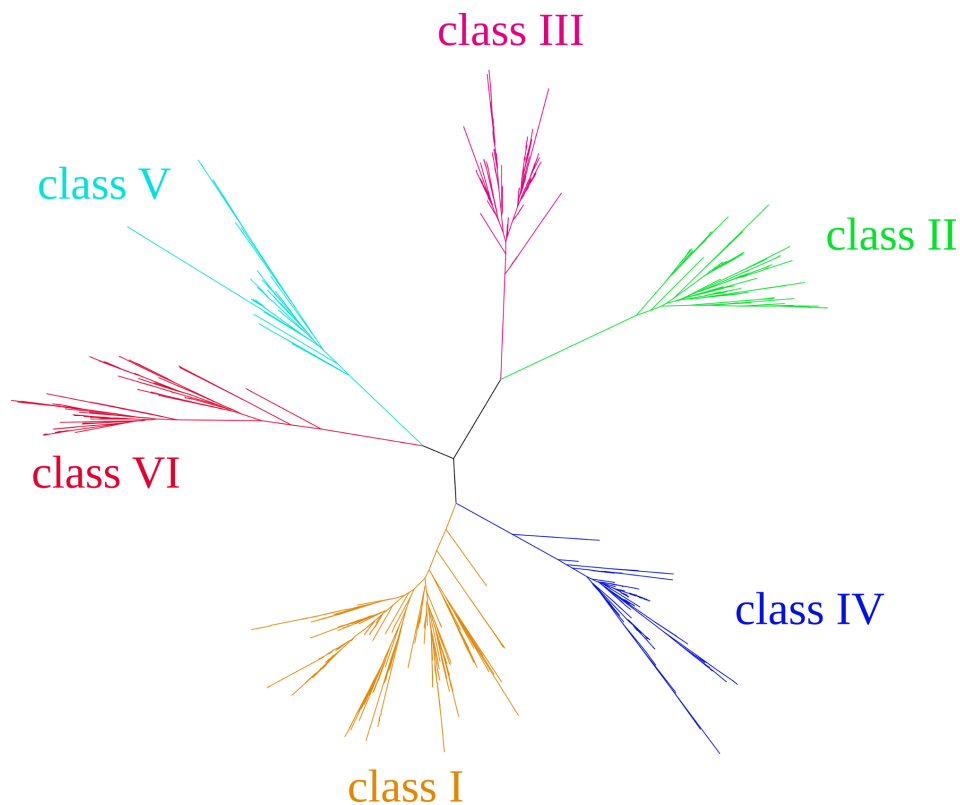
The aim of this project (paper II) was to summarise the mammalian ADHs currently available in the aforementioned sequence databases.

At the core of the project was the development of a pipeline that could be used to identify ADH members, confirm that they were correctly classified, and then add them to a local SQLite 3 database ([www.sqlite.org](http://www.sqlite.org)). A general visualisation of the pipeline is available in Figure 3.1, showing the use of BLAST<sup>60,61</sup> to identify sequences and HMMER<sup>65</sup> to confirm their classifications by matching the sequences against class-specific HMMs.



**Figure 3.1:** Visualisation of the pipeline used to identify and classify alcohol dehydrogenases available in the major sequence databases.

As the pipeline was designed to automate the whole process, the pipeline could be reused for later updates. However, the design turned out to be suboptimal for continuous updates, as it did not include any checks to confirm that sequences that had already been added were still present in the databases. The original analysis (performed during summer 2014) identified a total of 584 sequences belonging to 85 different mammalian species. A new run, performed



**Figure 3.2:** Phylogenetic tree covering the known mammalian ADH sequences as of March 2017.

in March 2017, identified 941 sequences from 145 mammalian species. Thus, the number of known mammalian ADH sequences keep increasing along with the number of species with at least one known member.

Both the original and new data show that the databases should be updated with a new class, as the phylogenetic tree clearly shows six classes, rather than the five that are usually noted in the literature (Figure 3.2). Many of the members of class VI ADH are currently misclassified as class V or “class I-like”.

The discovered sequences allowed further mapping of the mammalian ADHs. The number of ADHs in mammals varies between six in most non-primates to ten in vole, vole having four copies of ADH1 (ADH1:1–4), ADH2, ADH3, ADH4, ADH5, and two copies of ADH6 (ADH6A and ADH6B).

Three ADHs, namely ADH1, ADH2, and ADH6, have been duplicated in several species. Multiple copies of ADH1 are common in primates, but have also been observed in e.g. horses. Duplications of ADH2 have only been observed in hare animals, while ADH6 duplications have been observed in rodents.

Finally, many genomes still lack multiple expected ADH genes. The horse genome still lacks ADH2, ADH4, and ADH5, though ADH2 has been observed in Przewalski’s horse [NCBI: XP\_008508014.1]. The genes may have become pseudogenes, but their presence in related species imply that even genomes that are considered fully sequenced can still be improved.

### 3.3 CLASS V ALCOHOL DEHYDROGENASE

The class V alcohol dehydrogenase (ADH5), often called ADH6 in the sequence databases (Table 1.2), is a class of mammalian alcohol dehydrogenase with an as of yet unknown func-

tion. It was first identified using cross-hybridization with ADH1B cDNA probes in 1991<sup>79</sup>. The same protein was also identified as cDNA in deer mouse in 1993<sup>80</sup>. The human protein was investigated in 1991<sup>81</sup>, but as noted there, the data fits well with the “class II stomach ADH”, which correspond to the at that time not yet discovered ADH4.

Originally, ADH5 was believed to lack the last exon present in other ADHs, the ninth, due to a splice variant, but transcripts containing the exon has since been discovered<sup>82</sup>. In UniProt, the eight-exon variant is considered the main isoform with 368 amino acid residues, while the full version is annotated as isoform 2 with 375 amino acid residues.

The databases contain many misclassified ADH5 proteins, as most class VI ADHs are also annotated as ADH6, or even class V, sharing the annotation with most ADH5s in the sequence databases. As observed in section 3.2, while commonly located near the ADH5 branch in phylogenetic trees, the class VI ADHs clearly form a class of their own (Figure 3.2).

Being the last non-characterised ADH isoform that is present in humans, it is of interest to investigate its function and characteristics. The work presented in paper I and III covers an investigation of this protein, starting with in vitro approaches (paper I), and when the attempt to express and isolate the protein was unsuccessful, the focus shifted to computational methods (paper I and III).

### 3.3.1 In Vitro

Performing Northern blot analysis of rat ADH5 showed a strong signal in kidney, corresponding to both splice variants. The shorter variant was also observed in liver, stomach, duodenum, and colon. This differs from the expression data found in the Human Protein Atlas<sup>43</sup>, where the ADH6 entry notes high expression in liver, with kidney and the intestines having only low expression, similar to most other human ADHs.

The rat and human ADH5 proteins were expressed in *E. coli* and COS cells, respectively. The expression in *E. coli* did not generate any soluble human ADH5, but a fusion protein with glutathione-S-transferase (GST) could be expressed. However, this fusion protein did not show any activity with the traditional ADH substrates ethanol, octanol, benzyl alcohol, and hydroxymethylglutathione (HMGSH). In the case of rat ADH5 in COS cells, the protein could be expressed as a fusion protein with green fluorescent protein (GFP), but again with no observed activity for ethanol, octanol, or HMGSH. A similar fusion protein construct with GFP-ADH3 did have the expected activity, implying that the GFP construct should be functional.

In short, the native proteins could not be isolated, and fusion protein constructs did not have any activity with the traditional ADH substrates.

### 3.3.2 In Silico

As the in vitro experiments did not yield any isolated ADH5 proteins, the focus was shifted to analysing the sequences and structure using computational methods.

Initially (paper I), a model of rat ADH5 was generated using the MODELLER structure prediction method. The model was then analysed through a limited run of MD simulations using GROMACS, showing irregularities in the central  $\beta$ -sheets involved in the formation of dimers.

Continuing the analysis based on this observation, a model of human ADH5 was generated using a modified structure prediction method from the software ICM that was adapted to ADH analysis (paper III). The irregularities around the dimer-interacting region were observed in the human model as well.

Along with the structural models, the sequences of multiple ADH5 proteins were also investigated. At the time of the initial investigation (paper I), the number of ADH5 protein sequences in the databases was very limited, and there were only two mammals with at least one member in each of the six mammalian ADH classes. By the time of the continued investigation, the number had increased, with a total of 51 ADH5 sequences and ten species with at least one member per class.

By analysing the conservation of the positions within each of the six classes, multiple amino acid residues were identified as unique for the ADH5 proteins, including some positions with near-perfect conservation in ADH1–4 and 6, but with a different residue in ADH5, e.g. Gly305, replacing a Pro in most other mammalian ADHs. In total, ten unique amino acid residues were identified, with Lys51 being the only one that was conserved among all the ADH5 proteins. Lys51 is located at the active site, with the other ADHs having His, Tyr, or Thr at this position<sup>19</sup>.

Further, three residues, Leu295, Val299, and the aforementioned Gly305 were located in the dimer interaction region. Most classes of alcohol dehydrogenases have unique residues in this region to allow the formation of class-specific homodimers, but some positions are conserved between the classes as well, e.g. Gly305 which has a high level of conservation among all ADH1–4 enzymes.

Finally, the evolutionary pressure among the different classes was calculated from the nucleotide sequences using PAML. The pressure was expressed as the  $\frac{dN}{dS}$  quotient (section 2.3.3), where ADH5 was found to be the class with the weakest evolutionary pressure at 0.385, to be compared with the second-weakest pressure at 0.332 for ADH2 and the strongest at 0.126 for ADH3.

In summary, the problems with the expressions and activity of ADH5 were postulated to be due to irregularities in the dimer interaction region of ADH5, preventing the formation of dimers. As there is an evolutionary pressure, the protein should have some effect, but the results presented in paper I and III indicate that ADH5 does not form the traditional ADH dimers, and thus the title of paper III: “The odd sibling”.

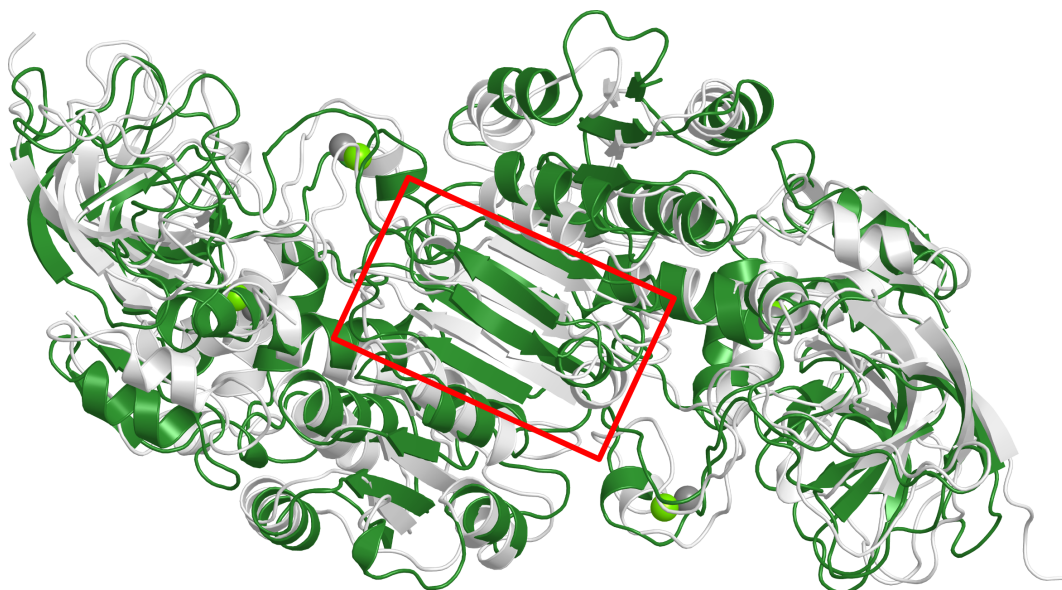
### 3.4 STUDYING THE MDR BINDING POCKET

Many MDR members have been characterised, but there are still several with unknown function. This project is aimed at defining the binding pocket, including both the coenzyme- and substrate-binding regions, and attempting to model the different types of active sites. To do this, nine MDRs that are present in humans; enoyl-[acyl-carrier-protein] reductase, mitochondrial (MECR)<sup>83</sup>, prostaglandin reductase 1–3 (PTGR1–3)<sup>20,84,85</sup>, quinone oxidoreductase (QOR)<sup>86</sup>, quinone oxidoreductase PIG3 (QORX)<sup>18</sup>, reticulon-4-interacting protein 1, mitochondrial (RT4I1)<sup>87</sup>, synaptic vesicle membrane protein VAT-1 homolog (VAT1)<sup>88</sup>, and synaptic vesicle membrane protein VAT-1 homolog-like (VAT1L), were analysed at the sequence and structure level. It should be noted that many of the proteins have changed names over time, and may thus be called other names in publications and databases. These include e.g. PTGR1 (leukotriene B4 12-hydroxydehydrogenase), PTGR2 and 3 (Zinc-binding alcohol dehydrogenase domain-containing protein 1 and 2), and MECR (mitochondrial 2-enoyl thioester reductase).

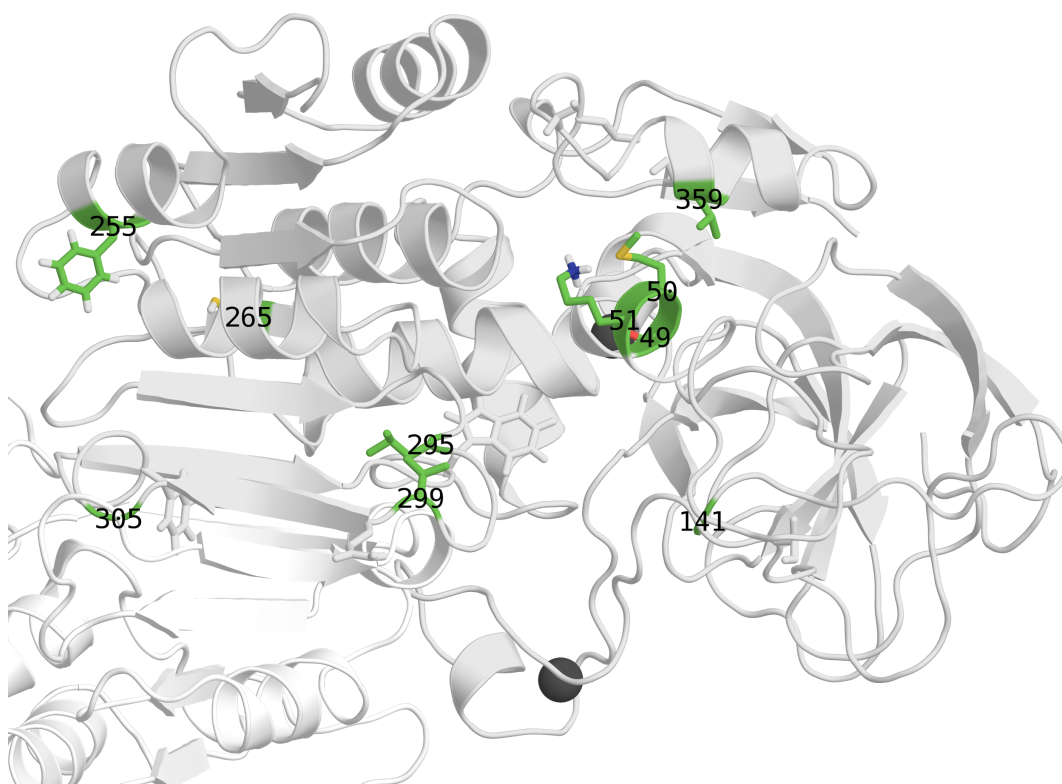
Eight of the nine proteins had structures available, while one did not.

The methodology included a large set of different computational approaches. In short, the conservation among the nine proteins was mapped to the structures, identifying positions that should be of importance for coenzyme as well as substrate binding.

The structure of the coenzyme-binding part of the binding pocket had a high level of



**Figure 3.3:** Comparison between structural models of human ADH1C and ADH5 after 20 ns of MD simulations. Note the differences in the central region where the dimer interaction occurs. Source: paper III



**Figure 3.4:** Unique positions in ADH5 mapped to a structural model of the same protein. Source: paper III

conservation over the different proteins, and the conserved residues were identified by combining the conservation data of all nine proteins. When the conserved residues were mapped to the residues being located near the coenzyme in the known structures, multiple important residues were identified. The list of residues contained many residues known to be involved in the interaction with the coenzyme, but also additional residues that may not be involved in the binding of the coenzyme, but rather in maintaining the structure so the relevant residues can interact.

The conservation in the substrate-binding part of the binding pocket showed larger variance, evident by the fact that the method of merging the conservation for all nine proteins did not yield any residues that are typically involved in substrate interactions.

Instead, the proteins had to be studied individually, and the conservation calculated only for that protein. This led to the identification of multiple important residues in each of the proteins. From this, three distinct types of substrate binding could be identified. The prostaglandin reductases use a combination of Tyr49 and Tyr245 (PTGR1 numbering), where Tyr245 was nearly completely conserved (>99%), while Tyr49 had a lower level of conservation (95%, 75%, and 99% for PTGR1–3 respectively). The other MDR proteins with known structure, with the exception of VAT1L, all had a highly (>95%) conserved Tyr in the substrate part. VAT1L did not have any conserved Tyr in the substrate part of the binding pocket, and the only conserved residues in the vicinity were Phe82 and Asp84, where Asp84 corresponds to a conserved Asp present in five of the other nine MDRs.

The MDRs could also be divided into four groups based on the structural localisation of the Tyr at position ~60. The first group would contain PTGR1 and PTGR2, the second PTGR3, QOR, QORX, MECR, and based on the sequence alignment, VAT1, and the final group RT4I1.

The conserved residues in the substrate binding pocket of the protein without known structure, VAT1, matched that of e.g. QORX.

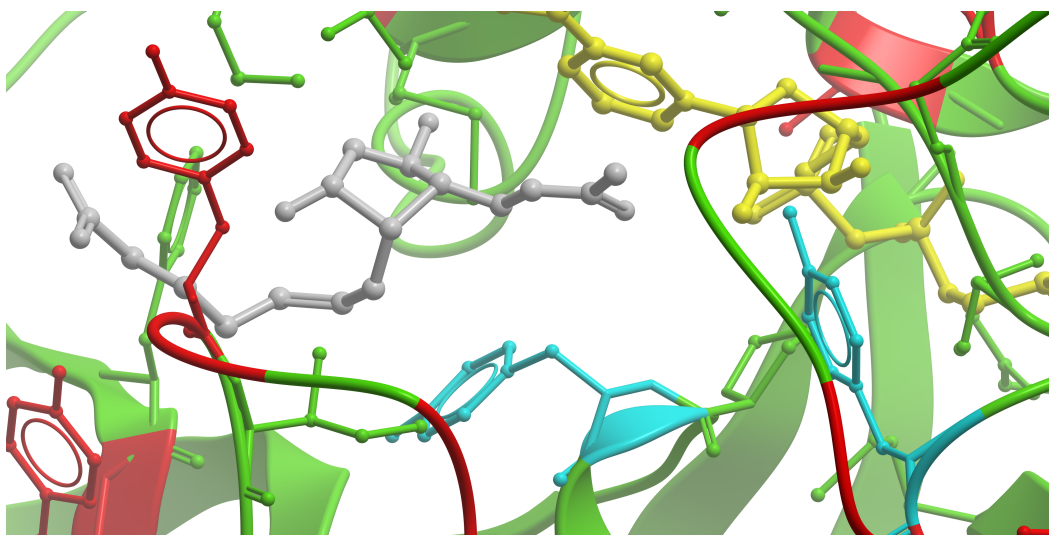
In the end, there were three unique substrate binding types identified among the nine MDRs; one in the PTGR proteins, one in the other MDRs except for VAT1L, and potentially one in VAT1L, as it did not have any conserved residues in the substrate-binding part that could be used to perform the traditional MDR reaction, indicating a third type of active site with an as of yet unknown function.

## 3.5 SUMMARY

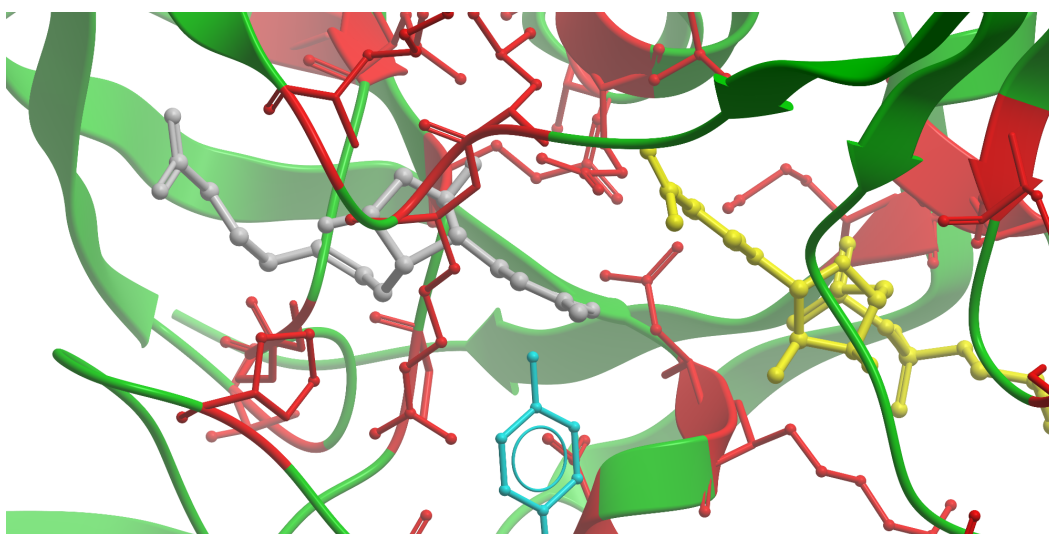
The work presented here covered the classification of mammalian ADHs, analysis of mammalian ADH5, and an analysis of the binding pockets of nine MDR members.

An automated strategy was developed to identify new mammalian ADHs that are available in especially the UniProt and NCBI protein databases. Along with this pipeline, an automatic classification system of ADH1–6 was also developed using HMMs. The pipeline was developed in 2014, and the classification is still accurate, as confirmed by phylogenetic trees. As expected, the number of ADH sequences in the databases keep increasing, and there are now 145 mammalian species with at least one known member. The pipeline can thus be used as an automated means of maintaining a database of known mammalian ADHs. It should also be noted that the pipeline allows the correct classification of ADH6, as opposed to the common misclassification of ADH6 proteins as ADH5 or ADH1-like in the sequence databases.

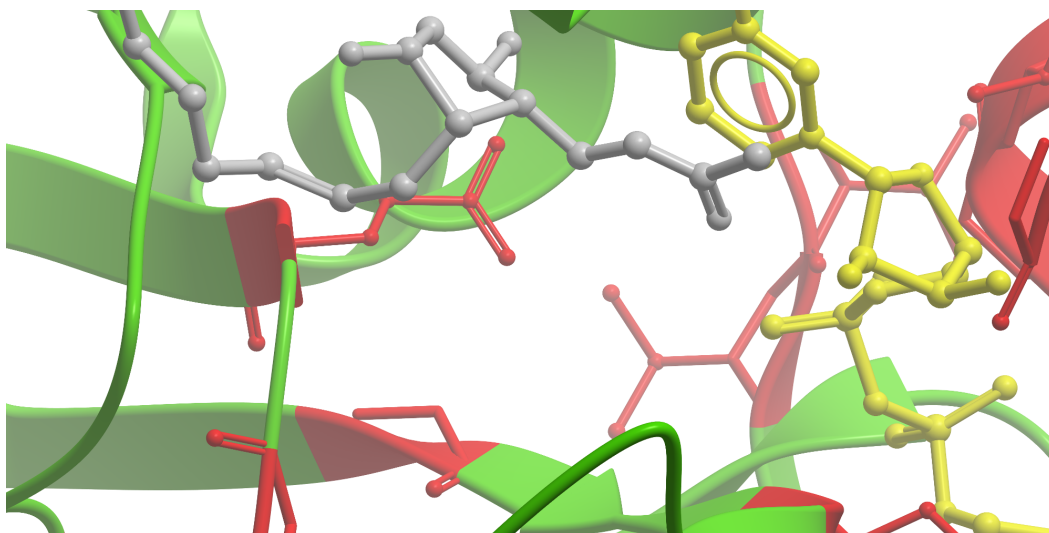
The class V alcohol dehydrogenase, ADH5, has still not been isolated, but the computational analysis imply that there are irregularities in the region for dimer formation, preventing the formation of dimers, which is known to be essential for the stability of mammalian



**PTGR2**



**QORX**



**VAT1L**

**Figure 3.5:** Substrate binding in the human MDRs. The three identified types use two Tyr (PTGR2; PDB id: 2ZB4), one Tyr (QORX; PDB id: 2J8Z), or neither (VAT1L; PDB id: 4A27). Yellow: NAD(P)H, Gray: substrate, Red: conserved residue, Cyan: Tyr. If a substrate or NAD(P)H was missing in the structure, the ones from PTGR2 were used as a reference.

ADHs. Still, there is an evolutionary pressure on the ADH5 genes to maintain their current sequences, and thus the protein should have some function. Taken together, these results imply that ADH5 does not form the traditional homodimers and has a function that differ from that of other ADHs.

The analysis of multiple MDR members showed that there are at least three different types of the substrate-binding part of the binding pockets. Of the proteins that were analysed, most seem to use Tyr to bind substrates, but there was also an example of a proteins that did not use Tyr (VAT1L), and also did not have any other conserved residues that were clear candidates for involvement in substrate binding.

Finally, the work presented here is an excellent example of how research is build on top of earlier work. The development of the pipeline used to identify and classify ADH members could be reused to identify ADH5 members, and with some slight modifications, the code could also be used to identify other MDR members. This is also true for the analysis developed to compare the classes of mammalian ADH while looking for unique residues could be extended and improved for the analysis of other MDRs.



## 4 GENERAL DISCUSSION

The ADH classes are originally defined based on the characteristics of ADH members that have been investigated experimentally<sup>19,23,25</sup>. As a result of this, ADH1–4 are clearly defined and correctly annotated in the databases. As ADH5 has never been isolated in its native form, the available information is very limited, and most “non-ADH1–4” sequences have been annotated as ADH5. By aligning and generating phylogenetic trees of the “ADH5” protein sequences, it has at multiple occasions, including in paper II, been shown that there is also a sixth class of mammalian ADH. This means that the ADH5 sequence data is easily contaminated by the ADH6 sequence data. Further, a large number of the mammalian ADH sequences that are present in the sequence databases often bear names similar to “uncharacterized protein” or “Alcohol dehydrogenase-like”, making it unclear what class they actually belong to (and if they are ADHs at all).

By developing an automated methodology for the retrieval and classification of ADH sequences, a complete set of all mammalian ADH protein sequences deposited in the major sequence databases could be obtained, and the set can also be kept up to date, being rerun whenever the databases are updated. The retrieved sequences can be used to map the presence of the ADH classes in different species, allowing observations such as the lack of ADH6 in primates, or that ADH5 is a pseudogene in mouse.

BLAST has been used to find homologs of sequences for a long time, including to find new ADHs. The main improvement with the new methodology is the automation, allowing the set of mammalian ADH proteins to be constantly up to date. In addition, the automated classification makes it unnecessary to confirm the classification of e.g. ADH5 proteins manually.

Further, it should be trivial to extend the methodology to cover more classes and other species by just adding new HMM definitions and determining appropriate cutoffs.

Human ADH5 is the only human ADH that has never been isolated and analysed as a native protein. The work presented here first attempted the isolation of the human and rat ADH5 proteins, but with little success. Fusion proteins were isolated, but they did not display any activity with any of the traditional ADH substrates. The lack of success implies that ADH5 differs greatly from the other human ADHs, as they can all be readily expressed and isolated as native proteins using the same techniques<sup>19</sup> that were used for ADH5.

The fact that ADH5 could be expressed as a fusion protein, but not as a native protein, implies that ADH5 may need a second protein to help stabilise it. The lack of success in isolating the native protein could then be explained by the fact that the needed protein is missing *in vitro*.

As the protein could not be isolated, the properties were investigated using computational methods instead, attempting to understand why the protein behaved so differently from the other ADHs, even though its sequence was similar. Based on the analysis, there were indications of irregularities in the region involved in the formation of dimers. Further, many highly conserved residues in ADH1–4 were replaced by other residues in ADH5. This could explain the lack of success with the isolation of the protein, as the formation of dimers is known to be important for the stability and function of the other human ADHs<sup>1, 19</sup>, fitting the experimental

data.

If the ADH5 protein is unstable and has no function, it would be expected that the gene should have many non-synonymous mutations, slowly turning the gene into a pseudogene. This is true in mice, where the ADH5 protein lacks multiple exons [NCBI Gene ID: 639769]. The human ADH5 also has known (albeit rare) mutations of residues that would render the other ADHs unusable. However, an analysis of the evolutionary pressure of ADH5 sequences from multiple species showed that there is no high fraction of non-synonymous mutations, most being synonymous instead. As such, it seems that there is an evolutionary pressure on the ADH5 protein sequence, forcing it to retain its current sequence.

The mouse genome contains two ADH6 genes, and it could be theorised that the function of the ADH5 gene has been overtaken by the one or both of the ADH6 genes. However, rat also has two ADH6 genes, as well as a complete ADH5 gene, making it improbable. There is the possibility that the rat ADH5 is being degraded, though inspection of the protein sequence gives no clear evidence of this, but as the properties of ADH5 are unknown it is difficult to evaluate it.

Thus, ADH5 seems to have a function, but it is different from that of other ADHs. It could be that it interacts with another protein *in vivo*, allowing to perform its effect, but it could also be functioning in a different environment *in vivo* than what was tested experimentally. The lack of activity could potentially be caused by Lys51, which may cause major changes at the active site, greatly altering the enzyme specificity from that of the other human ADHs, as it replaces a highly conserved His, Thr, or Tyr depending on class<sup>19</sup>. It seems that ADH5 should not be considered a pseudogene, but its function remains unknown.

There are currently 18 proteins identified as MDRs in humans. Of those, seven are the ADHs. Of the remaining eleven, one is SORD, similar to ADH, and one is the complex fatty acid synthase with multiple domains. The remaining nine were analysed at the sequence and structure level, focusing on their binding pockets. The important residues were identified using a combination of residue conservation and the localisation inside the binding pockets. As could be expected, the residues inside the coenzyme part of the binding pocket showed a high level of similarity, implying that all nine should be able to bind NADP(H)<sup>20</sup>.

The identified residues include most residues that are known to interact with the coenzyme, e.g. Lys178, Tyr193, and AGAVG at position 151–155 (most of the GXGXXG motif<sup>89</sup>), while e.g. Asn217 and Asn321 (PTGR1 residues/numbering) were missed. Apart from the known residues, the set also included some extra residues, e.g. Met124 and Pro125. These may be of importance for the structure of the coenzyme-binding part of the binding pocket, but do not interact directly with the coenzyme.

The developed methodology identified multiple residues in each of the proteins that were known to be involved in the substrate, but not all. In the case of QORX, only five out of eleven residues that were known to interact<sup>18</sup> were identified. In one case, this was due to inter-subunit interactions, the residue interacting with the substrate in the other chain, which could not be identified by the methodology, as it was developed to identify interactions only within a subunit. The remaining residues were mainly overlooked due to limited conservation of the residues in question. It does however mean that the residues that were identified had a greater importance for the general function of each enzyme, rather than the specific interactions needed for only a few species.

The substrate-interacting part of the binding pocket varied between the nine MDRs. Among the nine MDRs, three distinct types were identified. The PTGRs all had two Tyr located in the relevant part of the binding pocket, while the others (with the exception of VAT1L) all had one Tyr. VAT1L was the clear exception, with no conserved residues that are clearly involved in the substrate binding.

The methodology used to evaluate the MDRs with known structure was also used to

predict the residues involved in substrate binding in VAT1, showing that it seems to be an enzyme similar to that of QORX and MECR, with similar localisation of the Tyr.

The importance of the Tyr seems to vary between the different MDRs. In PTGR2, it is reported to be involved in the catalytic reaction, though it is not essential<sup>20</sup>. In QORX, it seems to help with the coordination of the substrate, but mutation of the residue may even increase the catalytic efficiency<sup>18</sup>. In the similar MECR protein, mutation of the Tyr causes a near complete loss of function<sup>83</sup>. Thus, the residue seems to have a large variation in importance, even when its localisation is near identical (as in the case of QORX and MECR), though the function may still be substrate coordination.

The use of Tyr in some MDRs is also similar to what is seen in many SDRs, e.g. the combination of Tyr152, Lys156, and Ser139 first seen in 3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase (P19992)<sup>90</sup>.

Thus, the human MDRs have four distinct mechanisms for substrate binding. The ADHs as well as SORD all have similar mechanisms, involving e.g. a  $Zn^{2+}$  ion. The PTGRs use two Tyr, VAT1L a non-defined mechanism, and the remaining MDRs a single Tyr. The involved MDRs are all homologs, but they still have different types of active sites.

The MDR superfamily is interesting from the aspect of evolution. The members vary considerably at the protein sequence level, but their folds are very similar. As noted, there are also multiple different substrate binding types as well as active sites. As an effect of this, the substrate specificity also varies largely. This means that neither the substrate specificity nor catalytic mechanism is conserved within the whole superfamily. Instead, the general fold has been conserved as it allows the binding of NAD(P)(H) as a coenzyme and, it seems, a good starting structure for performing catalysis of the conversion of alcohols and aldehydes. There are also cases where multiple genes have joined together, generating complex enzymes with multiple domains, e.g. fatty acid synthase<sup>21</sup>. Taken together, it is a strong example of the extent that evolution can change the properties of a single gene, over time generating a lot of different effects.

It is very difficult to define the ancestral gene of the MDRs. One suggestion would be ADH3 due to the effect on formaldehyde and/or S-nitrosoglutathione, already considered to be the ancestor of the ADHs<sup>31 19</sup>. This ADH is present in a very large array of species, both prokaryotes, eukaryotes, and archaea. However, there are many other MDRs that could also fill this role, and the variation at the active sites makes it very complex to evaluate the ancestry.

As has been shown in the work presented here, bioinformatics is a strong complement to experimental biochemical methods. The two fields has a symbiotic relationship, where biochemistry is required to feed the data to bioinformatics, and bioinformatics can in turn perform deeper analysis of the data, generating new hypotheses for the biochemical field.

The work performed on ADH5 is an example of this. The initial in vitro experiments showed that ADH5 could not be easily isolated, and its function could not be evaluated. The use of bioinformatics allowed the construction of a structural model and analysis of the sequence, which in turn led to the hypothesis that ADH5 is unstable due to being unable to form dimers. The addition of the evolutionary pressure analysis then complemented the hypothesis, showing that ADH5 should have some type of function. Thus the final hypothesis, ADH5 not forming dimers, but potentially interacting with something else. This hypothesis could then be tested in vitro, leading back to the field of biochemistry.

The medium-chain dehydrogenase/reductase superfamily is a superfamily that can be found in all phylogenetic branches, both prokaryotic and eukaryotic. The more than 170,000 sequences that are present in UniProt correspond to over 0.2% of all sequences included in the database. There are 20,199 human proteins in UniProtKB/Swiss-Prot, of which only 18 are MDRs, corresponding to 0.1%. Thus, it can be expected that at least 1 out of every 1000,

and potentially even more than 1 out of every 500, unique protein sequences belong to the MDR superfamily. The work presented here focused on specific MDRs, but it should be possible to extend the methods to cover more MDRs, and the analysis can give hints to the function of the thousands of MDR proteins that currently have unknown function.

## 5 ACKNOWLEDGEMENTS

First of all I would like to thank my supervisors Jan-Olov and Bengt for these valuable and helpful years, including all the help and knowledge they've given me. Mikko should also be included here, as without him I would never have started working with ADHs.

Hans should get a special mention for the collaborations and support.

Petter, Krillot, and Klara for always being ready with help, insults, useless information, or anything else that would or would not fit the situation. Gisela, Josefin, and Harriet for cheering me up when it was needed the most.

Michael, Frank, and Andreas, for being my main company for a long time. It was a sad time when you all finished.

Erik, thanks for all the great times and interesting discussions, and for being a friend since I first ventured into the field of biochemistry and molecular biology.

Anders, thank you for making me interested in molecular biology in the first place.

All the people from my time at MBB. Your help and company has been greatly appreciated. A special thank you goes to Alessandra, who always offered her help with any problems.

Everyone currently or previously at Gamma 6. This also includes all the people I interacted with from other parts of SciLifeLab. It was your presence that made my time here a pleasure.

И, наконец, спасибо Наталии и Лине за то, что делаете мою жизнь такой счастливой. Я вас люблю!



## 6 REFERENCES

1. Munroe R. xkcd [Webcomic]; 2017. Available from: <https://xkcd.com/1605/>.
2. Crick F. Central dogma of molecular biology. *Nature*. 1970 Aug;227(5258):561–563.
3. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. *PLoS Comput Biol*. 2011 Mar;7(3).
4. The International Human Genome Sequencing Consortium. International Consortium Completes Human Genome Project; 2003. Available from: <https://www.genome.gov/11006929/>.
5. Dayhoff MO, National Biomedical Research Foundation. Atlas of protein sequence and structure. vol. 1. Silver Spring [Md.]: National Biomedical Research Foundation; 1969. OCLC: 605459794.
6. Wu CH, Yeh LSL, Huang H, Arminski L, Castro-Alvear J, Chen Y, et al. The Protein Information Resource. *Nucleic Acids Res*. 2003 Jan;31(1):345–347.
7. Persson B, Zigler JS, Jörnvall H. A super-family of medium-chain dehydrogenases/reductases (MDR). Sub-lines including zeta-crystallin, alcohol and polyol dehydrogenases, quinone oxidoreductase enoyl reductases, VAT-1 and other proteins. *Eur J Biochem*. 1994 Nov;226(1):15–22.
8. Jörnvall H. Horse liver alcohol dehydrogenase. The primary structure of the protein chain of the ethanol-active isoenzyme. *Eur J Biochem*. 1970 Sep;16(1):25–40.
9. Persson B, Krook M, Jörnvall H. Characteristics of short-chain alcohol dehydrogenases and related enzymes. *Eur J Biochem*. 1991 Sep;200(2):537–543.
10. Hedlund J, Jörnvall H, Persson B. Subdivision of the MDR superfamily of medium-chain dehydrogenases/reductases through iterative hidden Markov model refinement. *BMC Bioinformatics*. 2010;11:534.
11. Persson B, Hedlund J, Jörnvall H. Medium- and short-chain dehydrogenase/reductase gene and protein families: the MDR superfamily. *Cell Mol Life Sci*. 2008;65(24):3879–3894.
12. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, et al. Pfam: The protein families database. *Nucl Acids Res*. 2014 Jan;42(D1):D222–D230.
13. Rao ST, Rossmann MG. Comparison of super-secondary structures in proteins. *J Mol Biol*. 1973 May;76(2):241–256.
14. Taneja B, Mande SC. Conserved structural features and sequence patterns in the GroES fold family. *Protein Eng*. 1999 Oct;12(10):815–818.

15. Stoop JM, Williamson JD, Conkling MA, Pharr DM. Purification of NAD-dependent mannitol dehydrogenase from celery suspension cultures. *Plant Physiol.* 1995 Jul;108(3):1219–1225.
16. Raj SB, Ramaswamy S, Plapp BV. Yeast Alcohol Dehydrogenase Structure and Catalysis. *Biochemistry.* 2014 Sep;53(36):5791–5803.
17. Jörnvall H, Persson M, Jeffery J. Alcohol and polyol dehydrogenases are both divided into two protein types, and structural properties cross-relate the different enzyme activities within each type. *Proc Natl Acad Sci U S A.* 1981 Jul;78(7):4226–4230.
18. Porté S, Valencia E, Yakovtseva EA, Borràs E, Shafqat N, Debreczeny JE, et al. Three-dimensional structure and enzymatic function of proapoptotic human p53-inducible quinone oxidoreductase PIG3. *J Biol Chem.* 2009 Jun;284(25):17194–17205.
19. Höög JO, Hedberg JJ, Strömberg P, Svensson S. Mammalian alcohol dehydrogenase - functional and structural implications. *J Biomed Sci.* 2001 Feb;8(1):71–76.
20. Wu YH, Ko TP, Guo RT, Hu SM, Chuang LM, Wang AHJ. Structural basis for catalytic and inhibitory mechanisms of human prostaglandin reductase PTGR2. *Structure.* 2008 Nov;16(11):1714–1723.
21. Jayakumar A, Tai MH, Huang WY, al Feel W, Hsu M, Abu-Elheiga L, et al. Human fatty acid synthase: properties and molecular cloning. *Proc Natl Acad Sci USA.* 1995 Sep;92(19):8695–8699.
22. Miinalainen IJ, Chen ZJ, Torkko JM, Pirilä PL, Sormunen RT, Bergmann U, et al. Characterization of 2-enoyl thioester reductase from mammals. An ortholog of YBR026p/MRF1'p of the yeast mitochondrial fatty acid synthesis type II. *J Biol Chem.* 2003 May;278(22):20154–20161.
23. Duester G, Farrés J, Felder MR, Holmes RS, Höög JO, Parés X, et al. Recommended nomenclature for the vertebrate alcohol dehydrogenase gene family. *Biochem Pharmacol.* 1999 Aug;58(3):389–395.
24. Jörnvall H, Landreh M, Östberg LJ. Alcohol dehydrogenase, SDR and MDR structural stages, present update and altered era. *Chem Biol Interact.* 2015 Jun;234:75–79.
25. Vallee BL, Bazzone TJ. Isozymes of human liver alcohol dehydrogenase. *Isozymes Curr Top Biol Med Res.* 1983;8:219–244.
26. Estonius M, Svensson S, Höög JO. Alcohol dehydrogenase in human tissues: localisation of transcripts coding for five classes of the enzyme. *FEBS Lett.* 1996 Nov;397(2-3):338–342.
27. Hellgren M, Strömberg P, Gallego O, Martras S, Farrés J, Persson B, et al. Alcohol dehydrogenase 2 is a major hepatic enzyme for human retinol metabolism. *Cell Mol Life Sci.* 2007 Feb;64(4):498–505.
28. Svensson S, Hedberg JJ, Höög JO. Structural and functional divergence of class II alcohol dehydrogenase. *European Journal of Biochemistry.* 1998 Jan;251(1-2):236–243.



29. Staab CA, Hellgren M, Höög JO. Medium- and short-chain dehydrogenase/reductase gene and protein families : Dual functions of alcohol dehydrogenase 3: implications with focus on formaldehyde dehydrogenase and S-nitrosoglutathione reductase activities. *Cell Mol Life Sci.* 2008 Dec;65(24):3950–3960.
30. Chou CF, Lai CL, Chang YC, Duester G, Yin SJ. Kinetic mechanism of human class IV alcohol dehydrogenase functioning as retinol dehydrogenase. *J Biol Chem.* 2002 Jul;277(28):25209–25216.
31. Danielsson O, Jörnvall H. "Enzymogenesis": classical liver alcohol dehydrogenase origin from the glutathione-dependent formaldehyde dehydrogenase line. *Proc Natl Acad Sci USA.* 1992 Oct;89(19):9247–9251.
32. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. *Genome Res.* 2002 Jun;12(6):996–1006.
33. Eklund H, Brändén CI, Jörnvall H. Structural comparisons of mammalian, yeast and bacillar alcohol dehydrogenases. *J Mol Biol.* 1976 Mar;102(1):61–73.
34. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci USA.* 1977 Dec;74(12):5350–5354.
35. Toribio AL, Alako B, Amid C, Cerdeño-Tarraga A, Clarke L, Cleland I, et al. European Nucleotide Archive in 2016. *Nucleic Acids Res.* 2017 Jan;45(D1):D32–D36.
36. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2016 Jan;44(Database issue):D7–D19.
37. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, et al. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.* 2002 Jan;30(1):27–30.
38. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. *Nucleic Acids Res.* 2017 Jan;45(D1):D635–D642.
39. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* 2015 Jan;43(Database issue):D204–D212.
40. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* 2017 Jan;45(D1):D190–D199.
41. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000 Jan;28(1):235–42.
42. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, et al. The PeptideAtlas project. *Nucl Acids Res.* 2006 Jan;34(suppl 1):D655–D658.
43. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. *Science.* 2015 Jan;347(6220):1260419.
44. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. *Nucleic Acids Res.* 2016 Jan;44(Database issue):D1202–D1213.

45. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992 Nov;89(22):10915–10919.
46. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970 Mar;48(3):443–453.
47. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981 Mar;147(1):195–197.
48. Chakraborty A, Bandyopadhyay S. FOGSAA: Fast Optimal Global Sequence Alignment Algorithm. *Scientific Reports*. 2013 Apr;3:1746.
49. Thompson JD, Koehl P, Ripp R, Poch O. BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*. 2005 Oct;61(1):127–136.
50. Pervez MT, Babar ME, Nadeem A, Aslam M, Awan AR, Aslam N, et al. Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evol Bioinform Online*. 2014;10:205–217.
51. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*. 2005 Feb;15(2):330–340.
52. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl Acids Res*. 2002;30(14):3059–3066.
53. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013 Apr;30(4):772–780.
54. Lassmann T, Sonnhammer ELL. Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. 2005 Dec;6:298.
55. Lassmann T, Frings O, Sonnhammer ELL. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res*. 2009 Feb;37(3):858–865.
56. Corpet F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*. 1988 Nov;16(22):10881–10890.
57. Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*. 2009 Jun;324(5934):1561–1564.
58. Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, et al. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst Biol*. 2012 Jan;61(1):90–106.
59. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. 2011 Jan;7(1):539.
60. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res*. 1997 Sep;25(17):3389–3402.

61. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009 Dec;10(1):421.
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990 Oct;215(3):403–410.
63. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science*. 1985 Mar;227(4693):1435–1441.
64. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. *Biol Direct*. 2012 Apr;7:12.
65. Eddy SR. Accelerated Profile HMM Searches. *PLOS Computational Biology*. 2011 Oct;7(10):e1002195.
66. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010 May;59(3):307–321.
67. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007 Aug;24(8):1586–1591.
68. RCSB PDB - Content Growth Report;. Available from: <http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100>.
69. Levinthal C. How to fold gracefully. *Mossbauer Spectroscopy in Biological Systems*. 1969;p. 22.
70. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One*. 2011 Dec;6(12).
71. Skwark MJ, Abdel-Rehim A, Elofsson A. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*. 2013 Jul;29(14):1815–1816.
72. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*. 1997 Apr;268(1):209–225.
73. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993 Dec;234(3):779–815.
74. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 2006 Jan;22(2):195–201.
75. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*. 2010;5(4):725–738.
76. Abagyan R, Batalov S, Cardozo T, Totrov M, Webber J, Zhou Y. Homology modeling with internal coordinate mechanics: deformation zone mapping and improvements of models via conformational search. *Proteins*. 1997;Suppl 1:29–37.
77. Paquet E, Viktor HL. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *BioMed Research International*. 2015 Feb;2015:e183918.

78. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015 Sep;1–2:19–25.
79. Yasunami M, Chen CS, Yoshida A. A human alcohol dehydrogenase gene (ADH6) encoding an additional class of isozyme. *Proc Natl Acad Sci U S A*. 1991 Sep;88(17):7610–7614.
80. Zheng YW, Bey M, Liu H, Felder MR. Molecular basis of the alcohol dehydrogenase-negative deer mouse. Evidence for deletion of the gene for class I enzyme and identification of a possible new enzyme class. *J Biol Chem*. 1993 Nov;268(33):24933–24939.
81. Chen CS, Yoshida A. Enzymatic properties of the protein encoded by newly cloned human alcohol dehydrogenase ADH6 gene. *Biochem Biophys Res Commun*. 1991 Dec;181(2):743–747.
82. Strömberg P, Höög JO. Human class V alcohol dehydrogenase (ADH5): A complex transcription unit generates C-terminal multiplicity. *Biochem Biophys Res Commun*. 2000 Nov;278(3):544–549.
83. Chen ZJ, Pudas R, Sharma S, Smart OS, Juffer AH, Hiltunen JK, et al. Structural Enzymological Studies of 2-Enoyl Thioester Reductase of the Human Mitochondrial FAS II Pathway: New Insights into Its Substrate Recognition Properties. *Journal of Molecular Biology*. 2008 Jun;379(4):830–844.
84. Yokomizo T, Ogawa Y, Uozumi N, Kume K, Izumi T, Shimizu T. cDNA cloning, expression, and mutagenesis study of leukotriene B<sub>4</sub> 12-hydroxydehydrogenase. *J Biol Chem*. 1996 Feb;271(5):2844–2850.
85. Yu YH, Chang YC, Su TH, Nong JY, Li CC, Chuang LM. Prostaglandin reductase-3 negatively modulates adipogenesis through regulation of PPAR $\gamma$  activity. *J Lipid Res*. 2013 Sep;54(9):2391–2399.
86. Fernández MR, Porté S, Crosas E, Barberà N, Farrés J, Biosca JA, et al. Human and yeast zeta-crystallins bind AU-rich elements in RNA. *Cell Mol Life Sci*. 2007 Jun;64(11):1419–1427.
87. Hu WH, Hausmann ON, Yan MS, Walters WM, Wong PKY, Bethea JR. Identification and characterization of a novel Nogo-interacting mitochondrial protein (NIMP). *J Neurochem*. 2002 Apr;81(1):36–45.
88. Koch J, Foekens J, Timmermans M, Fink W, Wirzbach A, Kramer MD, et al. Human VAT-1: a calcium-regulated activation marker of human epithelial cells. *Arch Dermatol Res*. 2003 Sep;295(5):203–210.
89. Wierenga RK, Terpstra P, Hol WGJ. Prediction of the occurrence of the ADP-binding  $\beta\alpha\beta$ -fold in proteins, using an amino acid sequence fingerprint. *Journal of Molecular Biology*. 1986 Jan;187(1):101–107.
90. Ghosh D, Wawrzak Z, Weeks CM, Duax WL, Erman M. The refined three-dimensional structure of 3  $\alpha$ ,20  $\beta$ -hydroxysteroid dehydrogenase and possible roles of the residues conserved in short-chain dehydrogenases. *Structure*. 1994 Jul;2(7):629–640.